DOCTOR OF PHILOSOPHY

Development and Application of Mass Spectrometry Based Methodologies to Study Ubiquitin-Like Modifiers

Kelly, Van

Award date:
2014

Development and Application of Mass Spectrometry Based Methodologies to Study Ubiquitin-Like Modifiers

Van Kelly

PhD, University of Dundee, December 2014

Table of contents

List of illustrations

List of abbreviations

BPC	base peak chromatogram
CID	collision-induced dissociation
DDA	data-dependent acquisition
DIA	data-independent acquisition
ESI	electrospray ionisation
FDR	false discovery rate
GO	gene ontology
HCD	higher-energy collisional dissociation
ID	identification
IEF	isoelectric focussing
LC	liquid chromatography
mcm5U	5-methoxycarbonylmethyluridine
MS	mass spectrometry
MS1	mass spectrum (survey scan)
MS/MS or MS2	tandem mass spectrometry (fragmentation spectrum)
MSn	mass spectrum, n>1
ncm5U	5-carbamoylmethyluridine
PMF	peptide mass fingerprinting
PSM	peptide spectrum match
RNA	Ribonucleic acid
s2U	2-thiouridine
SCX	strong cation exchange
SDS-PAGE	SDS poly acrylamide gel electrophoresis
SILAC	stable isotope labelling with amino acid in cell culture
TPP	Trans-Proteomic Pipeline
tRNA	transfer Ribonucleic acid
TUBEs	Tandem Ubiquitin Binding Entities
U34	Uridine-34
UBL	UBiquitin-Like protein (or UBiquitin-Like modifier)
wt	wild-type
XIC	extracted ion chromatogram

Acknowledgements

I would like to thank Dr Patrick Pedrioli for giving me this opportunity and for his guidance and patience. I also appreciate the guidance from my thesis committee, Arno Alpi, Geoff Barton, and Sarah McKim. To the whole Pedrioli Lab, Patrick, Kshitiz, Kamila, Todor and Cindy; thanks for all the coffee breaks. My experience in the lab would not have been as academically or socially fulfilling without you. Likewise to everyone in SCILLS and MRC, you have made my time here as a PhD student most memorable (and mostly memorable on JBC5!). I am grateful to those in SCILLS and MRC who donated samples to help beta-test my methodologies, and sometimes more than once - a particular thanks to Agne Kazlauskaite, Arno Alpi, and Thomas Gonatopoulos Pournatzis. A huge thank you to the cloning team for your expertise, I don't think anybody's research would go as smoothly without your support. And I can't imagine how much time the protein production team has saved me. Axel Knebel, you have purified … and re-purified to my satisfaction. Your efforts (and sarcasm) are beyond expectations.

Declarations

I declare that the following thesis is based on the results of investigations conducted by myself, and that this thesis is of my own composition. Work other than my own is clearly indicated in the text by reference to the relevant researchers or to their publications. This dissertation has not in whole, or in part, been previously submitted for a higher degree.

Van Kelly

I certify that Van Kelly has spent the equivalent of at least nine terms in research work at the School of Life Sciences, University of Dundee, and that he has fulfilled the conditions of the Ordinance General No 14 of the University of Dundee and is qualified to submit the accompanying thesis in application for the degree of Doctor of Philosophy.

Dr Patrick Pedrioli

Summary

Ubiquitin-like proteins (UBLs) have broad activities including modification of proteins, lipids, and tRNA. Analysis of both UBL conjugation sites and characterisation of system wide proteome responses is essential to understanding the impact UBLs have on regulating cellular systems. While mass spectrometry is a powerful analytical technique, identifying UBL isopeptides can be particularly challenging due their low abundance and the spectral complexity of UBL isopeptides with very long tryptic remnants. Furthermore, the complexity of whole proteomes delivers limitations to the throughput and depth of quantitative discovery proteomics. This thesis addresses some of the major technical challenges in UBL research. Biochemical and bioinformatic methodologies have been developed for the selective enrichment proteomics datasets, with a common theme of overcoming proteome complexity.

In the first instance, enrichment of diglycine isopeptides has been achieved through a novel method exploiting the polyglycine specificity of the bacterial protease-transpeptidase Sortase A (SrtA). Using a mutant with increased catalytic activity, SrtA mediates biotinylation of diglycine tryptic remnants, and also acts as a specific protease for release of isopeptides from streptavidin for analysis by mass spectrometry. This cost-effective approach to isopeptide enrichment is also applicable to linear N-terminal ubiquitylation. The method is demonstrated to offer greater than 100x enrichment and is exemplified on an in vitro ubiquitylation of MIRO1 by PINK1-activated PARKIN.

In contrast to a physical enrichment, a dataset enrichment of isopeptide identifications has been achieved through UBL isotope labelling. In vitro substrate modification with isotopically light and heavy UBL generates a characteristic isotopic doublet enabling isopeptides to be distinguished at the MS1 level. Candidate peptide identity can be assigned using high-resolution precursor mass and complementary MS2-level spectral interpretation with SUMmOn adds further confidence to isopeptide identities. Application to SUMO2 modification of putative substrate RNA guanine-7 methyltransferase (RNMT) revealed widespread SUMOylation at 16 different lysines by UBC9 despite lacking a consensus motif.

Finally, a quantitative proteomics workflow is presented that enriches whole proteome datasets by combining peptide identifications from unlabelled and SILAC proteomes. A software implementation, with additional tools for improved data quality management, is demonstrated to significantly improve proteome coverage and quantitative precision in unfractionated proteomes. An exemplification on elp3Δ/wt unfractionated yeast proteome reveals a subset of ELP-dependent uridine-34 tRNA modifications to be particularly important for efficient translation. Interestingly, the three mcm5s2U tRNAs which are co-modified by the UBL URM1, have a much greater impact on protein translation efficiency than the ELP-only modified mcm5U tRNAs.

Chapter I:
General Introduction

1.1 The ubiquitin family

1.1.1 Ubiquitin

Ubiquitin was discovered in 1975 and was recognised as being a ubiquitous polypeptide and a highly conserved protein present throughout eukaryotes (Goldstein et al., 1975). In almost 40 years since its discovery, ubiquitin has been well characterised as a key player in a vast number of cellular processes. The most characterised role is in protein recycling where ubiquitylation directs substrate proteins to the 26S proteasome for degradation. Protein half life is also regulated by the N-end rule mechanisms whereby a protein's N-terminus influences the degree of ubiquitylation and degradation (Varshavsky, 2011). Many misfolded proteins are recycled by ubiquitylation via E3-chaperone interactions (Kriegenburg et al., 2012). For instance, the endoplasmic reticulum associated protein degradation (ERAD) pathway retrotranslocates misfolded proteins from the ER back to the cytosol for degradation (Vembar and Brodsky, 2008). Crosstalk between the ubiquitin and autophagy pathways is also essential for cargo recognition and autophagic clearance of protein aggregates (Kraft et al., 2010). In contrast to bulk degradation mechanisms, the ubiquitin system can provide a highly specific targeting system. Through specific substrate recognition, a wide range of cellular processes are modulated, such as cell cycle, transcription, signal transduction, and development (Petroski and Deshaies, 2005). Furthermore, defects in the ubiquitin system are implicated in a wide range of human diseases including cancer and neurodegenerative diseases (Petroski, 2008).

Ubiquitin has a β-grasp fold and C-terminal glycine, which are common throughout the ubiquitin family. Ubiquitin functions by covalently attaching to substrate proteins which is mediated by a series of enzymes; an E1 activating enzyme, E2 conjugating enzyme, and E3 ligase. Ubiquitin can also be deconjugated from substrates by deubiquitylating enzymes (Figure 1.1). The conjugation reaction forms a covalent bond between the C-terminal carboxyl group of ubiquitin and an ε-amino group of a substrate lysine (Figure 1.2). Ubiquitin-like proteins (UBLs) are unique among post-translational modifications (PTMs) in that the modifying group is itself a protein, and the isopeptide bond formed is an amide bond with similar chemical properties to the peptide backbone. Ubiquitylation on cysteine, serine, and threonine residues have been reported, however their study has been limited due to the labile nature of esters and thioesters (Wang et al., 2012).

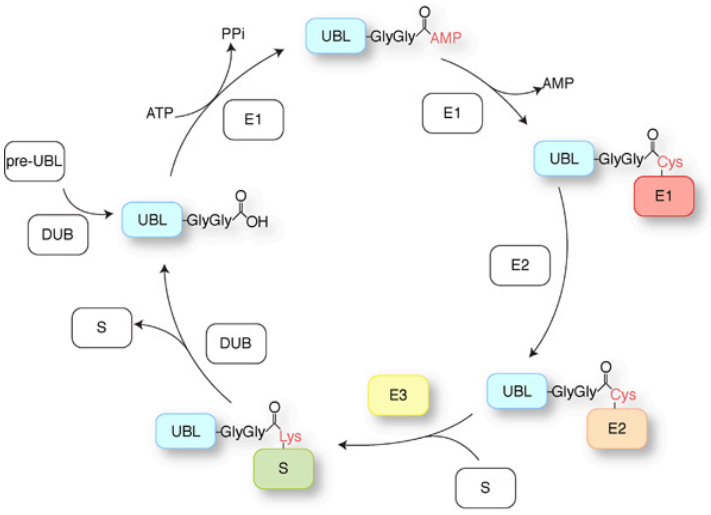


Figure 1.1 The ubiquitin conjugation pathway

Ubiquitin (Ub) is activated by an E1 enzyme to form a thioester intermediate, transferred to an E2 conjugating enzyme, also as a thioester intermediate, and then with the help from a substrate-recruiting E3 ligase, ubiquitin is transferred to a lysine on the substrate protein. Ubiquitin can also be deconjugated by a deubiquitylating enzyme (DUB) to reconstitute the original substrate and ubiquitin.

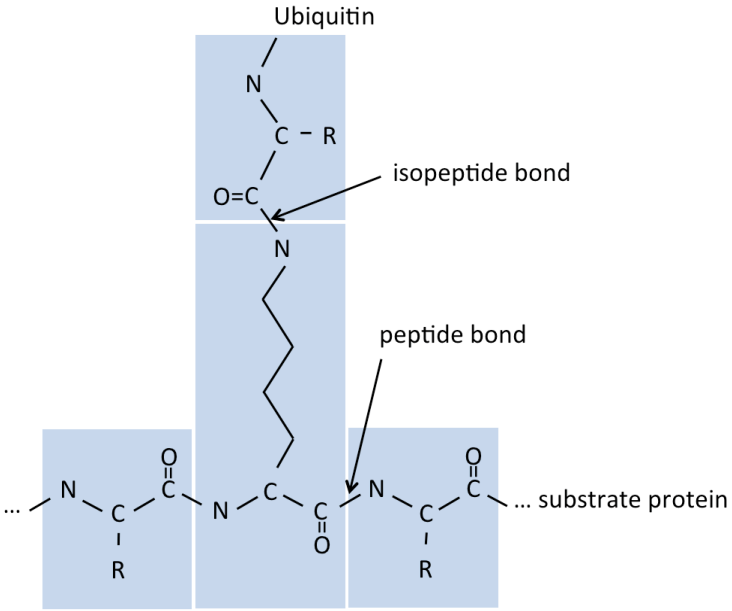


Figure 1.2 Ubiquitin is conjugated to a substrate lysine to form an isopeptide bond

Ubiquitin conjugation results in a covalent bond between the ubiquitin C-terminal carboxyl group and an ε-amino lysine on a substrate protein. Note that the isopeptide bond and the peptide bond are both amides and are chemically indistinguishable.

Ubiquitin can in fact be a substrate for itself by conjugating to one of its 7 lysines to form polyubiquitin chains. These chains can be homogenous, mixed, or even branched (Kirkpatrick et al., 2005). Ubiquitin chains can also form via the N-terminal α-amino group, which forms a typical peptide bond rather than an isopeptide. K48 chains and branched chains signal for degradation by the proteasome (Meyer and Rape, 2014), whereas monoubiquitylation and K63 linkages have non-proteolytic roles such as in DNA repair and protein sorting (Ikeda and Dikic, 2008; Lauwers et al., 2009; Spence et al., 1995). The function of K6, K11, K27, K29, and K33 ubiquitin chains are not well understood (Xu et al., 2009). The complexity of this modification becomes obvious when considering that a substrate can be monoubiquitylated, multi-monoubiquitylated, or modified with ubiquitin chains of various lengths and linkages.

Substrate lysines for ubiquitylation do not fall within a consensus motif (Xu et al., 2010). Instead, substrate specificity is mediated by more than 600 E3 ligases in the mammalian genome which act to recruit substrates and modulate E2 conjugation (David et al., 2011; Li et al., 2008). Most E3 ligases belong to the really interesting new gene (RING) ligase or related families and do not form a catalytic intermediate with ubiquitin. Cullin RING ligases are a multi-subunit family where the cullin acts as a scaffold for the formation of a highly dynamic E3 ligase (Mimura et al., 2010). The cullin complex itself is modified by another member of the UBL family, NEDD8. In contrast to the RING family of ligases, homologous to the E6AP carboxyl terminus (HECT) domain family of E3s do form a thioeseter intermediate with ubiquitin (Metzger et al., 2012).

1.1.2 NEDD8

NEDD8, or the yeast orthologue Rub1p, is a UBL with ~60% sequence identity to ubiquitin (Kumar et al., 1993). Like ubiquitin, NEDD8 is conjugated to target proteins in a similar manner but has its own set of E1, E2, and E3 enzymes. The primary target of NEDD8 is the cullin backbone of the cullin RING ligases, and the modification induces a conformational change to activate E3 ligase activity (Merlet et al., 2009). Cullin RING ligase regulation by NEDD8 exemplifies the interaction between members of the UBL family. *NEDD8* is essential in mammals, but surprisingly, *RUB1* is not essential in yeast (Liakopoulos et al., 1998). Although ubiquitin and NEDD8 have distinct enzymes mediating their conjugation, a high NEDD8 to ubiquitin ratio can results in NEDDylation via the ubiquitylation machinery (Hjerpe et al., 2012).

1.1.3 SUMO

There are 4 human small-ubiquitin-related modifier (SUMO) orthologues which share only a low sequence identity with ubiquitin (Figure 1.3). Despite the low primary sequence similarity they conform to the same β-grasp fold structure common to the ubiquitin family (Figure 1.4). The functionality of SUMO4 is questioned, although SUMO1-3 are conjugated by an E1 (SAE1/SAE2 dimer) and an E2 (UBC9) in a similar manner to ubiquitin (Wilkinson and Henley, 2010). Unlike ubiquitin however, conjugation via an E3 is not essential as UBC9 is sufficient to conjugate to lysines within a ΨKXE/D consensus motif, where Ψ is a large hydrophobic residue (Desterro et al., 1998; Rodriguez et al., 2001; Sampson et al., 2001; Sternsdorf et al., 1999). A reverse consensus motif and hydrophobic cluster SUMOylation motif have also been observed (Matic et al., 2010). SUMO2 and SUMO3 also contain a consensus motif at K11 permitting the formation of mixed SUMO2/3 chains. SUMO1 does not share the ability to form chains due to the absence of a consensus motif and prohibits further extension of SUMO chains *in vivo* (Matic et al., 2008). SUMO1 chains have however been observed *in vitro* (Cooper et al., 2005; Pedrioli et al., 2006).

```
SUMO1     MSDQEAKPSTEDLGDKKE-GEYIKLKVIGQDSSEIHFKVKMTTHLKKLKESYCQRQGVPMNSLRFLFEGQRIADNHTPKELGMEEEDVIEVYQEQTGG
SUMO2     MADEKPKE-----GVKTENNDHINLKVAGQDGSVVQFKIKRHTPLSKLMKAYCERQGLSMRQIRFRFDGQPINETDTPAQLEMEDEDTIDVFQQQTGG
SUMO3     MSEEKPKE-----GVKTE-NDHINLKVAGQDGSVVQFKIKRHTPLSKLMKAYCERQGLSMRQIRFRFDGQPINETDTPAQLEMEDEDTIDVFQQQTGG
ubiquitin M----------------------QIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG
          *                    :: *   .. : ::::   :.::    :::*:.  . *: * *: : :  *  :  ::.*..:.:  .  **
```

Figure 1.3 Sequence alignment between human SUMO paralogues and ubiquitin

Human SUMO2 and SUMO3 are almost identical (~97% identity) but differ significantly from SUMO1 (~50% identity). All SUMO proteins have very low similarity to ubiquitin (~20%). Multiple sequence alignment performed with CLUSTAL W.

Figure 1.4 Ubiquitin and SUMO2 share the β-grasp fold structure
Structural alignment between ubiquitin (PDB 2GBN, orange) and SUMO2 (PDB D207, blue) shows a very similar structure despite very low sequence similarity (~20% identity). Structures aligned using MacPyMOL

Many SUMO target proteins do not contains consensus motifs and require an additional level of specificity to be mediated by SUMO E3 ligases. Deletion of yeast SUMO E3 ligases, Siz1 and Siz2, significantly reduces protein modification by the yeast SUMO orthologue, Smt3 (Johnson and Gupta, 2001) indicating that E3-independent SUMOylation plays only a small role in yeast. SUMO E3 enzymes have been demonstrated to mediate non-consensus SUMOylation (Chiou et al., 2014) and enhance consensus site SUMOylation (Fuhs and Insel, 2011). Most SUMO E3s are thought to act to bring together UBC9 and substrates in a similar fashion to ubiquitin RING E3 ligases (Wilkinson and Henley, 2010). Although the SUMO modification is not a signal for proteasomal degradation, SUMOylation can indirectly induce protein degradation. The ubiquitin E3 ligase RNF4 recognises SUMO chains through multiple SUMO interaction motifs (SIMs) (Prudden et al., 2007; Sun et al., 2007; Uzunova et al., 2007; Xie et al., 2007). This results in substrate ubiquitylation and degradation of poly-

SUMOylated substrates (Tatham et al., 2008), thus demonstrating another example of interplay between UBL modifications.

1.1.4 URM1

Ubiquitin-related modifier-1 (*URM1*) was discovered in yeast and was observed conjugated to high molecular weight species in a UBA4 dependant fashion (Furukawa et al., 2000). Although the E2 and E3 enzymes remain undiscovered, the resemblances of Uba4p to ubiquitin's E1 and Urm1p to ubiquitin's β-grasp-fold structure place *URM1* as member of the UBL family. Urm1p conjugates to alkyl hydroperoxide reductase (Ahp1p) (Goehring et al., 2003) and to members of its own pathway (Van der Veen et al., 2011) under oxidative stress. However, protein urmylation has yet to be confirmed as a specific cellular mechanism as E2, E3, and deurmylating enzymes have not been identified.

Urm1p has also been recognised for its similarity to the bacterial thiamine synthesis pathway, in which a ThiF/ThiS acyldisulphide conjugate is required for thiazole moiety synthesis (Xi et al., 2001). Unlike the ubiquitin system, Urm1p was found to resemble the ThiF/ThiS conjugate by also forming an acyldisulphide intermediate with its E1, Uba4. Although Urm1p is not involved in thiamine synthesis, it acts as a sulphur carrier in the thiolation of tRNA (Figure 1.5). Urm1p therefore represents a unique member of the ubiquitin family in that it functions both as a sulphur carrier and a classic UBL and suggests an evolutionary link between the ubiquitin system and bacterial sulphur carriers.

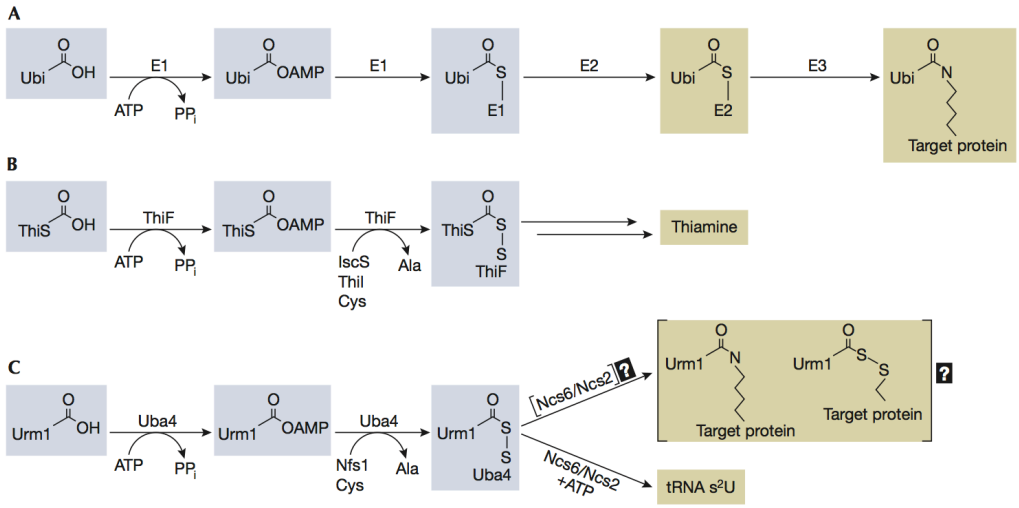


Figure 1.5 URM1 pathway resembles the bacterial thiamine synthesis pathway
The ubiquitin (A), ThiS (B), and Urm1 (A) pathways share only the first UBL adenylation of the carboxy terminus. The URM1 pathway subsequently resembles the initial steps of the bacterial thiamine synthesis pathway by formation of the acyldisulphide intermediate. Figure reproduced from (Pedrioli et al., 2008).

URM1's role as a sulphur carrier is essential for the thiolation of uridine-34 on tRNAs encoding for lysine, glutamine, and glutamic acid. The modified tRNA position 34 complements the third codon position of the three AAA, CAA , and GAA codons. Because tRNAs can 'wobble' at the third codon position (Crick, 1966), tRNA thiolation has the potential to alter the specificity of recognition. This can influence the extent of wobble to restrict recognition to the cognate codon or enhance recognition of near cognate codons (Agris, 2004).

The uridine-34 thiolation (s2) is also accompanied by an additional 5-methoxy-carbonyl-methyl (mcm5) modification (Figure 1.6) which is mediated by the Elongator protein complex (ELP) and each modification appears to affect formation of the other (Pedrioli et al., 2008). Deletion of *URM1* results in hypothiolation of tRNAs and lower tolerance to stress conditions such as starvation, oxidative stress, and temperature shifts (Goehring et al., 2003). Additionally, deletion of *URM1* results in translation defects

specifically for AAA, CAA, and GAA codons which results in a wide-spread impact on the proteome for genes rich in these three codons (Rezgui et al., 2013).

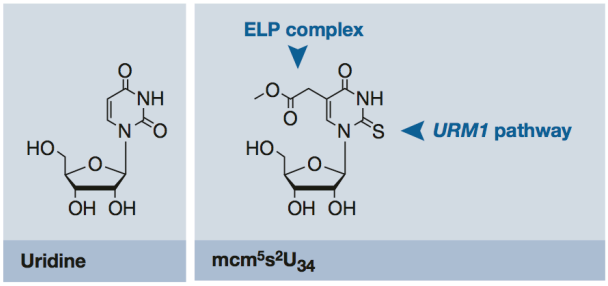


Figure 1.6 Chemical structures of uridine and mcm5s2-uridine
The uridine-34 of tRNAs recognising K, Q, and E codons are modified by both mcm5 and s2. Reproduced from (Pedrioli et al., 2008).

1.1.5 Other ubiquitin like proteins

Other members of the ubiquitin family exist including ISG15 (interferon stimulated gene 15kDa protein), FAT10 (human leukocyte antigen F-associated transcript 10), UFM1 (ubiquitin fold modifier 1), and ATG12 (autophagy-related protein 12), which conjugate to protein in a ubiquitin like manner. Other non-canonical members of the family include Hub1 (homologous to Ub), which lacks C-terminal glycine and is unlikely to conjugate to proteins. ATG8 (autophagy-related protein 8) is also most unusual in that it modifies lipids via a covalent bond to phosphatidylethanolamine (Van der Veen and Ploegh, 2012).

1.2 Introduction to mass spectrometry and proteomics

1.2.1 Mass spectrometry for proteomics

Bottom-up proteomics (also known as shotgun proteomics for complex samples) is currently the most conventional approach to protein identification (Figure 1.7). Proteins are first digested to generate a set of short peptides, typically using a high specificity

enzyme like trypsin. Tryptic peptides are separated by C18 chromatography for online analysis by mass spectrometry. As 'bottom-up' implies, the individual peptide identification need to be reassembled into their respective protein identities *in silico*. This workflow can also be conducted in combination with other offline fractionation techniques, including 1D and 2D SDS-PAGE, isoelectric focusing, or strong cation exchange.

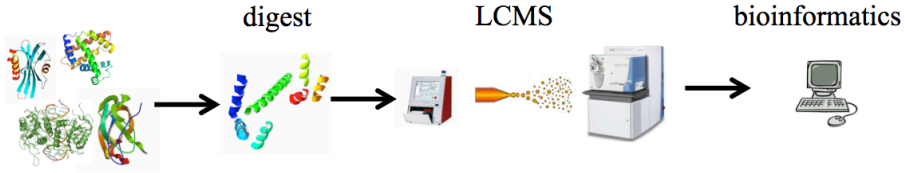


Figure 1.7 A typical bottom-up proteomics workflow
Protein samples are digested into peptides. Peptides are analysed by liquid chromatography, directly coupled to a mass spectrometer via electrospray ionisation. MS data is interpreted to assign peptide and protein identifications.

Soft ionisation techniques like electrospray ionisation (ESI) (Fenn et al., 1989) and matrix assisted laser desorption/ionization (MALDI) (Karas and Hillenkamp, 1988) enabled delivery of fragile biomolecules like proteins into the gas phase. Early MALDI techniques collected spectra on single protein digests and assignment of protein identity was conducted using only peptide masses by peptide mass fingerprinting (PMF) (Pappin et al., 1993). ESI permitted C18 reverse phase chromatographic separation of peptides to be coupled directly to a mass spectrometer. Mass spectrometers also gained the ability to isolate and fragment individual species, such as MALDI post-source decay and ion trap collision induced dissociation (CID). In combination with the ability to interpret peptide fragmentation spectra (MS/MS or MS2), LC-MS/MS made the analysis of more complex samples possible. Advancements in mass spectrometry sensitivity, speed, and resolution, has made mass spectrometry an indispensable tool for the analysis of complex proteomic samples (Clauser et al., 1999; Hebert et al., 2014).

While most mass spectrometry principles discussed are generally applicable to any high resolution mass spectrometer, specific applications refer to the Thermo Velos Orbitrap mass spectrometer (Figure 1.8) which was the instrument used during this research. The Velos Orbitrap mass spectrometer is a hybrid instrument, meaning that it is comprised of two mass analysers, an ion trap and an orbitrap. The ion trap is a low mass resolution analyser, but it is relatively fast and sensitive. The ion trap performs ion isolation and fragmentation (MS2), and can pass ions forward to the orbitrap analyser. The ion trap is also unique in that it can perform multiple rounds of isolation to further fragment an MS2 fragment (MS3). The Orbitrap mass analyser is an electrostatic trap which detects ion oscillations and converts them to mass spectra using Fourier transform (Makarov, 2000). The orbitrap resolution is proportional to transient time and although it can record very high resolution mass spectra it is slow compared to the ion trap. Hybrid instruments also have the unique ability to operate both analysers simultaneously enabling rapid collection of ion trap spectra in parallel during an orbitrap scan.

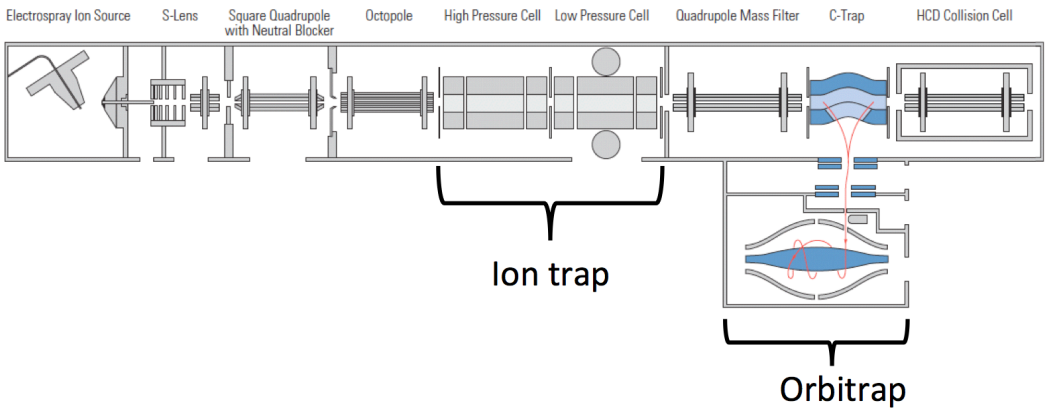


Figure 1.8 Schematic of the hybrid Velos Orbitrap mass spectrometer

The main features of this hybrid mass spectrometer are the inclusion of two independent mass analysers, the ion trap and the orbitrap. A high-energy collisional dissociation (HCD) cell is also available for an alternative peptide fragmentation technique.

1.2.2 LCMS data structure and modes of acquisition

LCMS data can be considered as regular samplings over a continuous multidimensional domain (Figure 1.9). As peptides elute from C18 chromatography, they are observed with isotopic resolution over the m/z domain, with an intensity, and elution time range. This data is often reduced to represent a feature as a discrete triple of m/z, intensity, and elution time. Peptide fragmentation data is also an additional discontinuous level of information associated with peptide features (Figure 1.10). Typically, a 'high-low' strategy is taken where MS1 is collected at high resolution and the acquisition cycle is completed with one or more MS2 data acquisitions at low resolution. This strategy ensures high quality precursor mass accuracy and rapid MS2 acquisition.

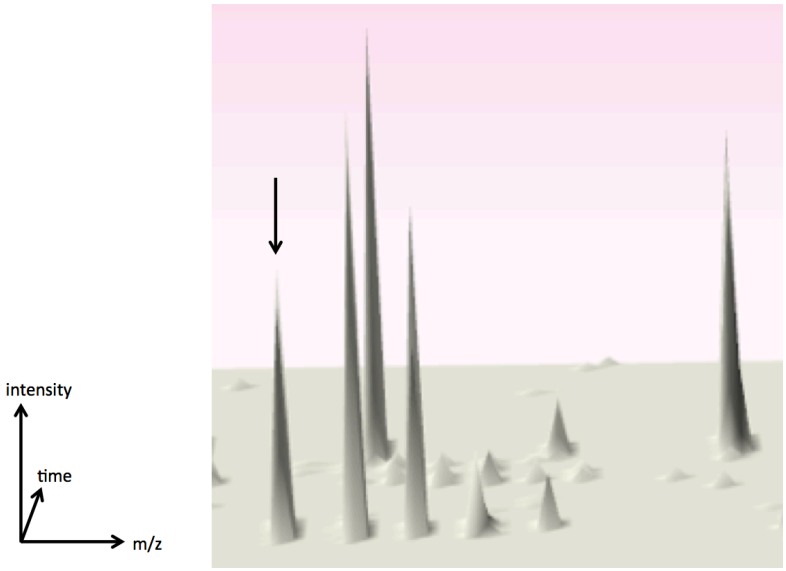


Figure 1.9 A peptide is observed as a 'feature' in 3-dimensional LCMS data landscape

A peptide feature is observed as multiple isotope peaks over the m/z, intensity, and time domains. A feature can be deconvoluted from continuous 3D data to discrete m/z, intensity, and elution time of its monoisotopic peak (indicated with arrow). A feature may also have associated MS2 spectra.

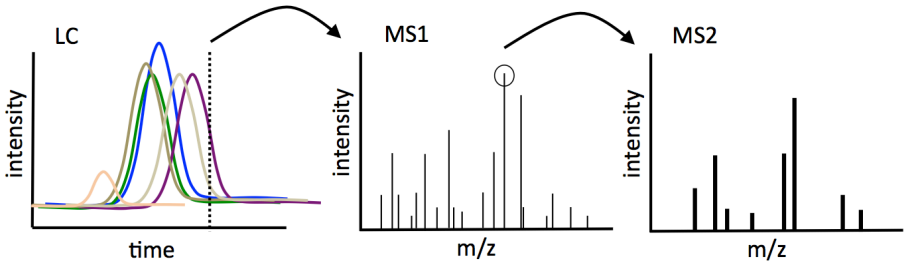


Figure 1.10 A depiction of LC-MS/MS data-dependent acquisition sequence
During LC separation of peptides, MS1 scans survey the eluting peptides. One or more ions are selected for fragmentation to generate MS2 scans before repeating the cycle for the duration of the LC gradient. Note that an additional MS3 event (not shown) is possible using ion traps but is not routinely done in proteomics.

The MS1-level feature landscape of a sample can be analysed by continuously collecting MS1 scans. In contrast, the addition fragmentation data must be acquired with some decision logic. For the analysis of samples with unknown composition, the duty cycle is usually directed by data-dependent acquisition (DDA). In DDA, the decision on what precursor ions to select for fragmentation is determined by the observed ions in the preceding MS1 scan (usually the most abundant n peaks). Hybrid instrument can operate two detectors for MS1 and MS2 acquisition simultaneously, but to do so the decision logic must come from quick average-resolution 'preview' scan (Figure 1.11). For a more targeted data acquisition, an inclusion list containing precursor m/z and time can be used to guide selection of ions to fragment. A targeted analysis can still be a DDA analysis, in which case precursors on the list will be acquired if observed in the MS1 scan. In data independent acquisition (DIA), MS2 spectra are acquired without any consideration for observed MS1 signals and therefore relies heavily on scheduling MS2 event times.

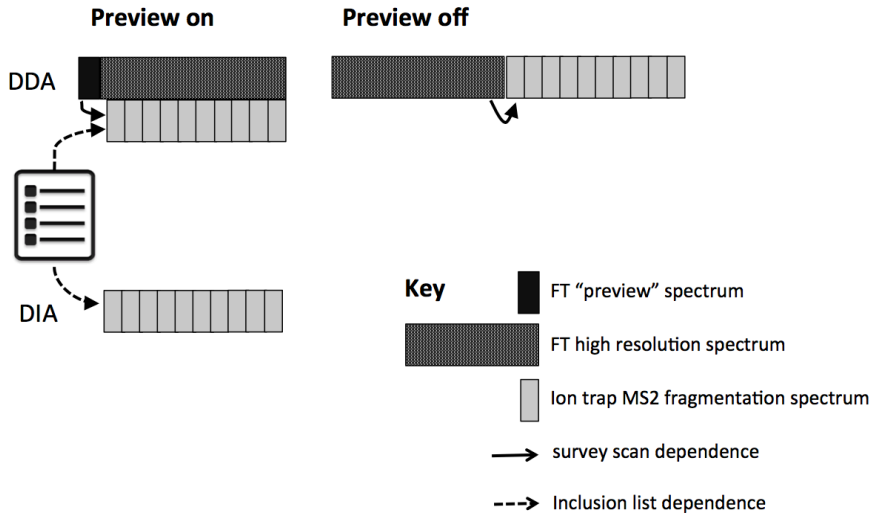


Figure 1.11 Illustration of alternative MS2 data acquisition methods

Hybrid orbitap instruments normally acquire a 'preview' survey scan to enable data-dependent acquisition (DDA) of MS2 spectra to proceed simultaneously to a high resolution MS1. In non-hybrid instruments, or where preview is not used, the high resolution MS1 scan acts as a survey which delays DDA decision logic. An inclusion list can guide precursor selection in DDA analysis or schedule MS2 events in data independent analyses (DIA).

1.2.3 Computational proteomics

Identification of peptides and proteins is relatively straightforward for sequenced genomes. A protein database can be digested *in silico* to generate peptides and their theoretical spectra. Candidate precursor ion identities can be determined through accurate mass matches to theoretical peptide masses. In simple mixtures and high mass precision, peptide mass fingerprinting (PMF) is sufficient to identify proteins (Pappin et al., 1993), although MS2 spectral comparison in the normal approach with MS2 capable mass spectrometers. The theoretical and experimental spectra can be compared to generate a similarity scores. Higher precision on both precursor and fragment ions will minimise search space and improve specificity. Many software platforms are capable of such searches including SEQUEST (Eng et al., 1994), Mascot (Perkins et al., 1999), and XTandem (Craig and Beavis, 2004). Concatenating forward and reverse databases can

help determine true identifications from false positives using statistical models, such as PeptideProphet in the Trans-Proteomic Pipeline (TPP) (Keller et al., 2002).

1.2.4 Quantitative proteomics

Comparing proteomes for quantitative differences can give insight into molecular cause and consequence of different biological states. Mass spectrometry allows us to both identify and quantify peptides simultaneously and numerous methods for quantitative comparison have been developed including label-free, iTRAQ, spectral counting, and stable isotope labelling with amino acids in cell culture (SILAC) (Wasinger et al., 2013). When isotope labelling during cell growth is a possible experimental approach (Figure 1.12), SILAC offers favourable quantitative precision (Oppermann et al., 2013). Lysine and arginine auxotrophic cells are grown in light media or media containing isotopically heavy arginine and lysine. Mixing cells before sample processing minimises handling variation and digestion with trypsin ensures that most peptides contain an isotopic label. Ratios between light and heavy isotope labels for identified peptides can be assembled into protein ratios using many supporting software solutions including XPRESS (Han et al., 2001).

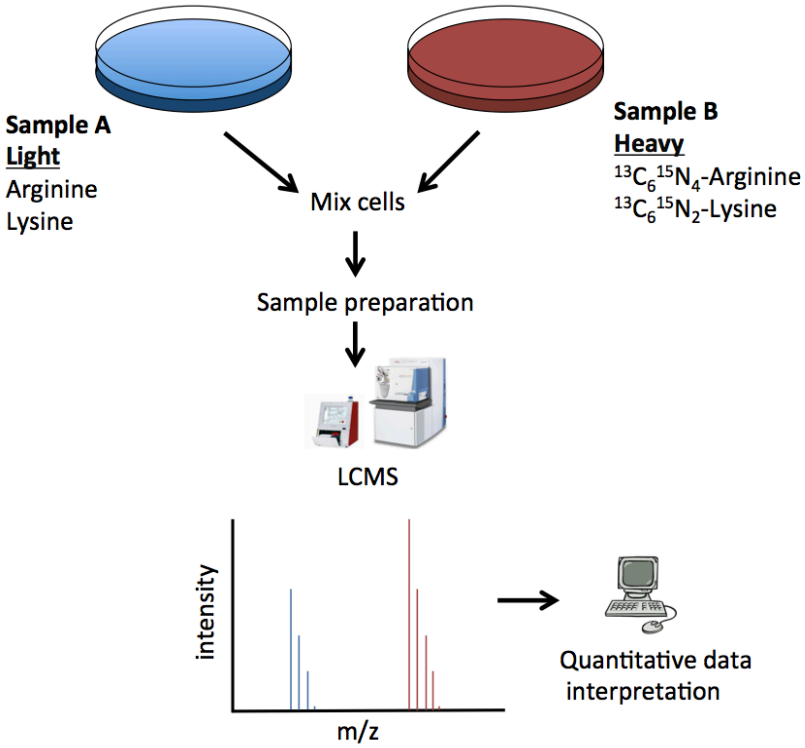


Figure 1.12 Overview of SILAC labelling

Cell cultures are grown in light or heavy stable isotope labelled arginine and lysine. Cells are mixed prior to samples processing including tryptic digestions, followed by LCMS analysis. Computation analysis derived relative quantitation on identified peptides and proteins.

1.3 Mass spectrometry of ubiquitin-like modifiers

The proteomics field faces numerous challenges in characterising UBL modifications. One of these challenges is overcoming the low abundance of UBL modified proteins. Lessons from other PTMs, such as phosphorylation, has taught us that peptide-level enrichment techniques are critical to enable efficient PTM analysis by mass spectrometry. Enrichment techniques like IMAC and TiO_2 have enabled global proteome analysis of phosphopeptides, however, their differing enrichment specificities highlight the need for multiple enrichment strategies (Bodenmiller et al., 2007). Ubiquitin has recently received much attention in global proteome experiments due the development of two K-GG specific antibodies. They too have overlapping but non-

identical substrates suggesting that further enrichment strategies could also benefit ubiquitin studies (Wagner et al., 2012).

Other members of the ubiquitin family have yet to receive a similar level of attention, partly due to a lack of peptide level enrichment strategies. Even in samples where enrichment is not necessary, mass spectrometry analysis of UBL conjugation sites is further challenged by the complex nature of the modification. Unlike small modification like phosphorylation and acetylation, UBLs leave a peptide remnant on modified lysines complicating their analysis (Jeram et al., 2009).

While many of the challenges are specific to detection of UBL modification sites, the effect UBLs have on the proteome as a whole should not be overlooked. Quantitative analysis of the entire proteome can reveal potential UBL substrates by perturbing members of ubiquitin degradation pathways (Chen et al., 2012). Furthermore, global proteome analyses can also reveal UBL-dependent dynamic changes in signalling networks (Bennett et al., 2010). Achieving a wide proteome coverage with sensitive quantitation in whole proteomes is still a developing field (Colaert et al., 2011).

1.4 Thesis roadmap

Each of the three results chapters present research addressing some of the main challenges in UBL analysis. Each chapter is a semi-independent body of work and approaches UBL research from the bench, at the mass spectrometer, and *in silico*.

Chapter II, "Sortase Mediated Biotinylation for Isopeptide Enrichment (SoMBIE)", addresses the low abundance challenge and presents a novel method for the enrichment of ubiquitin isopeptides.

Chapter III, "Isotope Coded Isopeptide Detection (ICID)", addresses the challenges in confident identification of isopeptides. In this chapter, techniques are presented for the detection and identification of isopeptides in LCMS analyses, with a focus on UBLs with long tryptic remnants.

Chapter IV, "HyperProphet", presents a new workflow and software to improve coverage and accuracy in quantitative SILAC analyses of unfractionated proteomes.

Chapter II
Sortase Mediated Biotinylation for Isopeptide Enrichment (SoMBIE)

2.1 Introduction

2.1.1 Ubiquitylation substrate identification

The identification of ubiquitylated substrates is important for understanding biological systems and their points of regulation. This is often achieved by perturbation of the ubiquitin system, such as mutating or deleting members of the ubiquitylation machinery. In such cases, targets for ubiquitin dependent degradation can be indirectly inferred. Upon deletion of an E3 ligase, up regulated proteins are revealed as a potential substrate for the E3 due to their accumulation. For example, a label-free quantitative proteomics approach has successfully identified filamin A and filamin B as substrates of ASB2, which is a subunit responsible for substrate recruitment to the cullin-RING E3 ligase complex (Burande et al., 2009). While this is a straightforward approach for the discovery of candidate ubiquitin substrates, this approach is limited to degraded substrates rather than non-degredatory roles of ubiquitylation. Furthermore, many candidates could be observed due to indirect effects. Proteomic investigations into the role of the ubiquitin system responses to disease states or exogenous stimuli requires a more direct focus on ubiquitylated substrates to distinguish between ubiquitin dependent responses and other indirect proteome changes. It is therefore important to focus specifically on the ubiquitylated subset of the proteome. This can be achieved by enriching for ubiquitylated substrates prior to proteomic analysis.

2.1.2 Protein-level enrichment of ubiquitylated substrates

A major limitation in the identification of ubiquitin substrates is the low abundance within the proteome. While most proteins will be ubiquitylated at some point in their life cycle, the pool of ubiquitylated substrates in the proteome is small due to rapid degradation once modified. Only a small portion of a protein may be modified or only modified during specific cell stages. Enrichment of ubiquitin, and with it the conjugate substrates, is a common approach to dealing with the dynamic range challenge. In genetically tractable systems, N-terminal fusions between affinity tags and ubiquitin offer a powerful way to enrich for ubiquitin. Note that C-terminal tags are not possible due to the requirement for the free C-terminal carboxyl group involved in conjugation.

A range of tags have proven successful, commonly 6HIS for its tolerance to denaturing conditions (Peng et al., 2003; Tsirigotis et al., 2001), but also more complex double affinity tags such as 6HIS–Myc (Hitchcock et al., 2003) and Strep-HA (Danielsen et al., 2011). For cells where genetic manipulation is impractical, ubiquitin purification can be achieved using the FK2 antibody (Matsumoto et al., 2005). Ubiquitin binding domains are also available for ubiquitin affinity purification, such as ubiquitin-associated (UBA) domain-containing proteins Rad23 and Dsk2 (Mayor et al., 2005). Construction of a tandem-repeated ubiquitin-binding entities (TUBEs), based on UBA domains of ubiquilin 1, increases affinity for polyubiquitin to enhance enrichment of polyubiquitylated substrates (Hjerpe et al., 2009; Shi et al., 2011a).

2.1.3 Ubiquitylation substrate lysine identification

Enrichment for ubiquitylated substrates increases the specificity for assigning ubiquitin substrate candidates. However, non-specific binding to affinity media or co-precipitation of interactors increases false assignments. Identifying the specific lysine

to which ubiquitin is conjugated is an important step in validating substrates as genuine targets. Knowledge of the conjugated lysine also enables downstream investigation into the biological significance of target lysines through mutagenesis. When dealing with individual substrates, mutagenesis can be used as an initial approach but is often impractical due to the number of lysines, and the modified lysine is not necessarily specific within a protein zone of modification (Mattiroli and Sixma, 2014). Alternatively, determining the specific lysine that is modified can be achieved by mass spectrometry based isopeptide identification.

As for a typical proteomics workflow, the most common approach for identifying ubiquitin isopeptides is by tryptic digestion followed by mass spectrometry. As the C-terminal sequence of ubiquitin (…RLRGG) contains tryptic sites near the C-terminal carboxyl conjugation site, the tryptic remnant of ubiquitin is a very short GG modification, or sometimes a short LRGG missed cleavage. Figure 2.1 compares the fragmentation of a linear peptide to an equivalent ubiquitin isopetpides. Because the tryptic remnant is a small modification, fragmentation within this group does not contribute significantly to the overall spectrum.

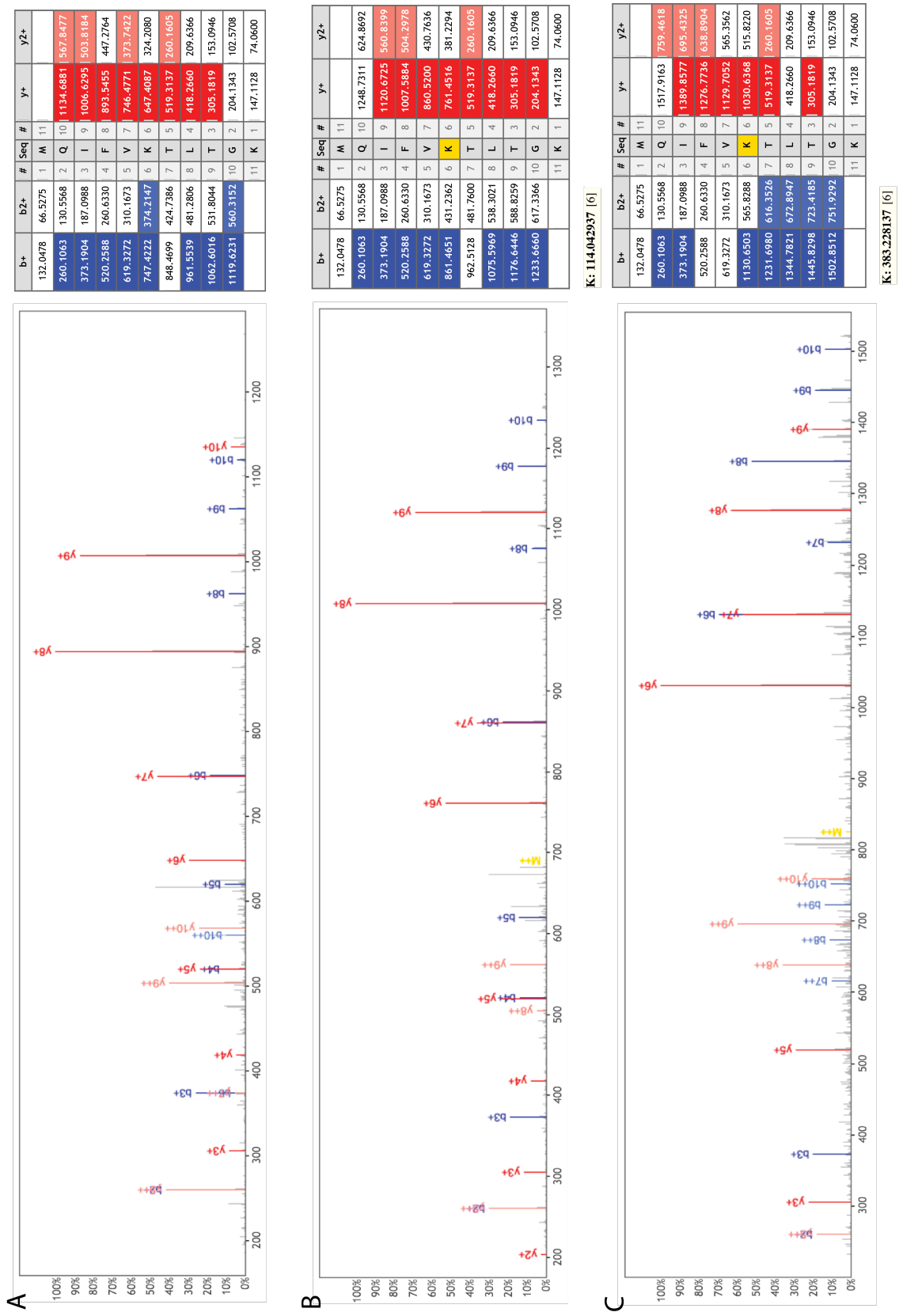


Figure 2.1 A comparison of peptide spectra with ubiquitin tryptic remnants

Spectrum identifications for ubiquitin peptides encompassing the K6 lysine for the missed cleaved unmodified (A), GG modified (B) and LRGG modified (C) peptides. A shift in the modified lysine residue mass is observed allowing confident identified but without interference from the ubiquitin tryptic remnant fragment ions. Data was acquired from a tryptic digest of K6 ubiquitin dimer.

From an *in silico* standpoint, the modified K-GG residue can be considered as a single entity with a +114.04292 Da mass, thus resembling simple modifications like an acetylation or methylation. An LRGG remnant giving a +383.22809 Da modification can also occur due to tryptic missed cleavage of the ubiquitin side chain and should be accounted for during database searches. Over-alkylation with iodoacetamide may cause lysine double alkylation resulting in a modification of identical atomic composition to a diglycine modification (Nielsen et al., 2008). Care should be taken to avoid this by using alternative alkylating reagents with differing modification mass or less reactive reagents such as chloroacetamide. Addition tryptic missed cleavages should be permitted because the modified residue is no longer efficiently cleaved by trypsin, although cleavage of the modified K-GG residue has been reported (Denis et al., 2007). Validation of results should also attempt to distinguish between a modification with GG and semitryptic cleavage after asparagine (N) which has a near identical mass of 114.04293 Da, although this will only occur with semi-tryptic database searches permitting. Management of false positive modified peptides should be considered carefully as global target-decoy results do not necessarily hold true for the modified subset of a search result, particularly when combining results from multiple search engines (Shi et al., 2011b). But on the whole, if isopeptide spectra can be acquired, assignment of ubiquitylated peptides identities to the spectra is not a limiting factor.

Even in sample enriched for ubiquitylated proteins, there are typically many candidate proteins identified that do not have isopeptides assigned. Note that identifying a protein may be straight forward as only one of the multiple available tryptic peptide are required to be identified. In contrast, identifying an individual peptide or isopeptide of interest may be more challenging due to low signal intensity or

poor fragmentation. For direct purification of substrate proteins (as opposed to ubiquitin enrichments) and *in vitro* reactions, modification stoichiometry can be low making isopeptide identification difficult.

2.1.4 Peptide-level enrichment of ubiquitylated substrates

Overcoming isopeptide detection challenges can be achieved by direct purification of the isopeptides. Unlike protein level purifications, these methods focus on the unique structure of isopeptides that differentiate them from typical peptides. Upon tryptic cleavage, ubiquitin modified peptides acquire two N-termini; the native peptide N-termini and a second N-termini derived from the ubiquitin remnant. This unique feature enables isopeptides to be targeted for enrichment using amine reactive fluorous affinity tags (Brittain et al., 2005). In this approach, lysine reactivity was blocked by guanidination of peptides with O-methylisourea and N-terminal amines were derivatised with an N-hydroxysuccinimide ester of 1H,1H,2H,2H- perfluorohexanoic acid. Peptides with two flourous tags bind more stringently to fluorous-functionalised silica allowing singly modified peptide to wash away in moderate solvent washes. While this is an intriguing approach, it has been demonstrated only on digests of polyubiquitin but has not been shown to be effective in more challenging samples. Furthermore, isopeptide amines remain blocked with the uncharged perfluorohexanoate modification. These peptides were effectively analyses by MALDI, but the modification is likely to be detrimental to analysis by ESI.

Another unique feature of ubiquitin isopeptides is the unique epitope generated by the K-GG tryptic remnant. This epitope was exploited by generating an antibody with specificity to the internal K-GG and without cross reactivity to N-terminal diglycine (Xu et al., 2010). While initial results were modest, subsequent optimised usage

(Wagner et al., 2011), and development of a second antibody (Kim et al., 2011) has resulted in purification of ubiquitin isopeptides from whole cell digests with impressive identification rates being reported. Each antibody has an overlapping but non-identical sequence preference around the modified lysine suggesting a sequence bias for one or both antibodies (Wagner et al., 2012). This suggests there may be a class of isopeptides that are poorly enriched using these antibodies. This technique has now become the dominant approach for the enrichment of ubiquitin isopeptides for analysis by mass spectrometry.

The available methods for enrichment of ubiquitin isopeptides are limited. By nature of the methods, having specificity towards internal KGG epitope, or requiring two N-termini, neither method permits enrichment of N-terminal ubiquitylation. The method presented in this chapter is a novel approach to enrichment of ubiquitin isopeptides which also permits targeting N-terminal diglycine modifications. In this approach, a method is developed to take advantage of the polyglycine specificity of the enzyme Sortase A.

2.1.5 Sortase A anchors virulence factors on gram positive bacterial cell wall

Staphylococcus aureus is major cause of human infection and as a major health concern, the study of *S. aureus* virulence is an active area of research. The virulence factor Protein A is well known for its immunoglobulin binding properties, and its expression on the cell surface contribute to *S. aureus* pathogenicity (Patel et al., 1987). A host of cell surface proteins are presented on the surface of *S. aureus* and other gram positive bacteria that containing an LPXTG motif (Marraffini et al., 2006; Navarre and Schneewind, 1999). In a screen for cell wall sorting mutants, sortase A (henceforth just

referred to as sortase) was identified as a potential mechanism for sorting these cell wall proteins (Mazmanian et al., 1999). *S. aureus* Sortase A is the type specimen for a large collection of sortase enzymes that have been discovered throughout gram positive bacteria, some of which have differing sequence specificities (Dramsi et al., 2005). Further characterisation confirmed that sortase was indeed the enzyme responsible and that cysteine-184 was necessary for enzymatic activity suggesting a transpeptidation was occurring (Ton-That et al., 1999). Sortase recognises the LPXTG sorting signal motif and proteolytically cleaves at the threonyl-glycyl amide bond to form a thioester intermediate. In a second transpeptidation step, the substrate is transferred to the N-terminus of pentaglycine of the cell wall peptidoglycan (Figure 2.2). The discovery and characterisation of sortase has lead to the introduction of new enzyme in the toolkit for biotechnological applications.

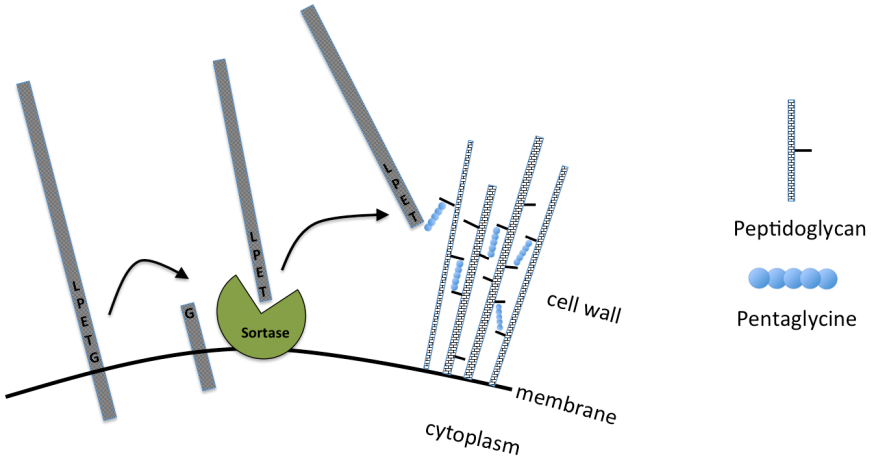


Figure 2.2 Sortase conjugates bacterial cell surface proteins to peptidoglycan

Sortase recognises the LPXTG cell sorting motif, then via a thioester intermediate, conjugates the C-terminal threonine to the N-terminus of pentaglycine within the outer cell wall peptidoglycan.

2.1.6 Current use of Sortase A in biotechnology

Sortase has been used in biotechnological applications for its transpeptidase activity. More importantly, the specificity in which it recognises its substrates can be exploited to direct sortase to conjugate two approriately labelled biomolecules (Figure 2.3). Fusion of two proteins can be achieved by incorporating the LPXTG motif near the C-terminus of one protein, and polyglycine at the N-terminus of another. This application of sortase is a useful tool for *in vitro* ligation between protein domains where gene fusions fail (Levary et al., 2011). Conjugation reactions have also been demonstrated for creating cyclic polypeptides and glycopeptides (Wu et al., 2011) and cyclic cytokines (Popp et al., 2011). Application of sortase also extend to conjugations to other organic molecules where polyglycine is modified with a wide range of molecules such as fatty acids, fluorescent probes, biotin, nucleic acids, sugars, and antibiotics (Proft, 2010).

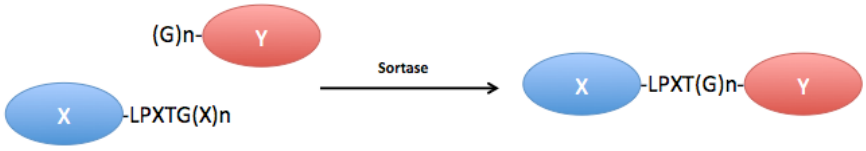


Figure 2.3 Sortase can be exploited to mediate conjugation between biomolecules

By incorporation the LPXTG motif near the C-terminus of protein X and polyglycine at the N-terminus of protein Y, a conjugated X-Y product can be generated. Proteins can be substituted for alternative reagents and solid supports as long as the LPXTG and polyglycine are accessible to sortase.

2.1.7 Sortase A specificity

The utility of sortase lies largely in its specificity. Extended reaction times with variant peptide substrates have revealed a small tolerance for individual amino acid substitutions for the LPXTG motif, with only the glycine being absolutely specific

(Kruger et al., 2004). However, this tolerance was only observed in the absence of the preferred substrate. Non-specific reactivity has otherwise not been reported and sortase is considered specific for the LPXTG motif. Other sortase variants from *S. aureus*, and sortases from other species have different specificities, but should not be confused with *S. aureus* Sortase A.

Although sortase recognises pentaglycine as the second substrate in its native scenario, sortase has been confirmed as more tolerant to shorter glycine nucleophilic substrates. *In vitro* kinetic analysis revealed that the length of the polyglycine is only required to be a diglycine and extending to three or more glycines only slightly improves reactivity (Huang et al., 2003). No reactivity was detected against a single glycine and GV and GA dipeptides were found to be possible substrates for sortase but with much higher Km values. Huang et al. (2003) concluded that the nucleophile binding site of sortase is specific for diglycine. Diglycine was also reported as favoured substrate for peptide-peptide conjugations (Pritz et al., 2007) and for attachment to solid supports (Chan et al., 2007) but with reduced recognition of substrates where the second glycine was substituted for other amino acids.

2.1.8 Objectives of research

Limited peptide level purifications options are available for ubiquitin isopeptides. Also, no option available is applicable towards linear ubiquitin linear peptides. Sortase mediated conjugation has been identified as a potential mechanism for targeting digycine motifs in tryptic peptides and, although sortase has been used in a wide range of applications, the applicability towards ubiquitin isopeptide enrichment has not been investigated to date. Because isopeptides posses an unusual conformation there is a possible hindrance to recognition between sortase and the ubiquitin tryptic glygly

remnants giving uncertainty to this application. The degree to which non-specific reactivity would interfere was also uncertain as there are reports of GX reactivity at low level relative to the preferred GG substrate. The objectives of this research was to determine the feasibility of using sortase, and if feasible, develop a methodology whereby peptides containing a diglycine ubiquitin tryptic remnant are enriched for mass spectrometry analysis. Although this chapter has focused on purification of ubiquitin isopeptides, the same principles apply to all glygly modifications, including those from NEDD8 and ISG15, which also have a glygly tryptic remnant.

This chapter presents work that is predominantly a methods development project which has gone through a number of modifications since its conception. The initial experimental design will be presented along with key adaptations made during the development of this method, followed by example applications of Sortase Mediated Biotinylation for Isopeptide Enrichment (SoMBIE).

2.2 Materials and Methods

2.2.1 Peptides

All were custom syntheses at >90% purity, except GG which was available at >99.5%):

GG (Sigma)

LIFAGK(GG)GLEDGR (JPT Peptide Technologies GmbH)

biotin-VELPRTGEE (Pepceuticals Ltd, UK)

biotin-VELPKTGEE (Generon Ltd, UK)

biotin-AGGGGG-VELPKTGEE (JPT Peptide Technologies GmbH)

biotin-Ttds-VELPKTGEE (JPT Peptide Technologies GmbH)

biotin-Ttds-SGGSGGSGGGGGVELPKTGEE (JPT Peptide Technologies GmbH)

biotin-SGGSGGSGGGGGVELPKTGEE (JPT Peptide Technologies GmbH)

biotin-SGGSGGSGGGGGVELPKTGEEHHHHHH (JPT Peptide Technologies GmbH)

biotin-SGGSGGSGGGGGVELPETGEEHHHHHH (Thermo Fisher Scientific GmbH)

2.2.2 Protein reagents

Protein for *in vitro* ubiquitylation were as published (Kazlauskaite et al., 2014)):

Flag-ubiquitin (Boston Biochem), MDYKDDDDKGG precedes ubiquitin M1

Human recombinant Ube1, purified from Sf21 insect cell line

Human His-SUMO-Parkin, expressed in *E. coli*, tag cleaved off by His-SENP1

MBP-TcPINK1 expressed in *E. coli,* (and a kinase-inactive version, D359A)

Human recombinant UbcH7, expressed in *E. coli*.

6His-Sumo-Miro1, expressed in *E. coli*.

6HIS-HALO-TUBEs (Tandem-repeated Ubiquitin-Binding Entities), expressed in *E. Coli*. Sequence based on Genebank Q9UMX0, expressed in pET28a vector (pET28a-6HIS-TEV-HALO-4xUbiquilin)

Sortase (wt and variants): Sortase 60-206 protein sequences are preceded by 6HIS and an S-tag (underlined), followed immediately by sortase A Q60

Sortase A, wt (MRC-PPU cloning Ref: SC20983)

MGSSHHHHHHSSGLVPRGSHMKETAAAKFERQHMDSQAKPQIPKDKSKVA
GYIEIPDADIKEPVYPGPATPEQLNRGVSFAEENESLDDQNISIAGHTFIDRPNYQ
FTNLKAAKKGSMVYFKVGNETRKYKMTSIRDVKPTDVEVLDEQKGKDKQLTL
ITCDDYNEKTGVWEKRKIFVATEVK

Sortase A, 3x point mutant, Tag deletion ΔG2, D160N K190E K196T (MRC-PPU cloning Ref: DU22269)

MSSHHHHHHSSGLVPRGSHMKETAAAKFERQHMDSQAKPQIPKDKSKVAG
YIEIPDADIKEPVYPGPATPEQLNRGVSFAEENESLDDQNISIAGHTFIDRPNYQF
TNLKAAKKGSMVYFKVGNETRKYKMTSIRNVKPTDVEVLDEQKGKDKQLTLI
TCDDYNEETGVWETRKIFVATEVK

Sortase A, 5x point mutant, Tag deletion ΔG2, P94R D160N D165A K190E K196T (MRC-PPU cloning Ref: DU22272)

MSSHHHHHHSSGLVPRGSHMKETAAAKFERQHMDSQAKPQIPKDKSKVAG
YIEIPDADIKEPVYPGPATREQLNRGVSFAEENESLDDQNISIAGHTFIDRPNYQF
TNLKAAKKGSMVYFKVGNETRKYKMTSIRNVKPTAVEVLDEQKGKDKQLTLI
TCDDYNEETGVWETRKIFVATEVK

2.2.3 Sortase cloning and expression

All coning was performed by the cloning services team (Dundee MRC-PPU). Expression of Sortase A was performed by the Protein Production and Assay Development (Dundee MRC-PPU). The sortase A (gene ID ADC38675) was cloned from *Staphylococcus aureus* genomic DNA. The sequence corresponding to amino acids 60-206 was inserted into a pET28a plasmid including a HIS6 and S-tag. pET28a-

HIS6-S-tag-SrtA60-206 was transformed and expressed in *Escherichia coli* BL21 cells. Cells were grown in LB medium containing 50 μg/mL kanamycin (Sigma) at 37 °C to an OD600 of 0.8. 250 μM isopropyl 1-thio-β-D-galactopyranoside (Sigma) was added to induce expression and cell were harvested by centrifugation at 3,000 x *g* for 5 min after 16 hr growth at 15 °C. Cells were lysed by sonication in 25 mM HEPES/NaOH pH 7.5, 150 mM NaCl, 0.1 mM EGTA, 0.1 mM EDTA, 1 mM Leupeptin (Sigma), 5 mM imidazole. Lysate was clarified by centrifugation at 17,000 x *g* for 5 min and applied to Ni Sepharose Fast Flow (GE Healthcare) and eluted in 25 mM HEPES/NaOH pH 7.5, 100 mM NaCl, 0.1 mM EGTA, 0.03% Brij-35, 400 mM imidazole, 10% glycerol and buffer exchanged into 50 mM HEPES pH 7.5, 10% glycerol, 150 mM NaCl, 1 mM Dithiothreitol (DTT)(Formedium). Sortase point mutants were derived from the above plasmid and were expressed as outlined above.

2.2.4 Re-purification of sortase

Additional purification of the Sortase 3x point mutant was conducted by size exclusion on a HiLoad 16/600 Superdex 75 column (GE Healthcare) in 50 mM Tris pH 7.5, 150 mM NaCl, with protein signal detection at 280 nm. 5 mg of protein was fractionated and the most abundant fraction was re-concentrated on NiNTA agarose (Qiagen), eluted in 50 mM Tris pH 7.5, 150 mM NaCl, 400 mM imidazole, then dialysed overnight against 50 mM Tris pH 7.5, 150 mM NaCl using a 1 kDa dialysis unit (Amersham).

2.2.5 SDS-PAGE and western blotting

A fixed proportion of 1/1000 of the sample was collected during purification and analysed by 15% SDS-PAGE with coomassie staining. A duplicate gel was transferred to Hybond-C Extra Nitrocellulose membrane (Amersham) and blocked in 4%

BSA/TBS/0.1% Tween 20 overnight. Incubation with 1/1000 FK2 anti-ubiquitin antibody (Millipore) was followed by 1/3000 Goat Anti-Mouse IgG (H + L)-HRP Conjugate in 1% BSA/TBS/0.1% Tween 20 for 1 hr. Incubations were separated and followed by five 15 min washed in TBS/Tween 20 and developed using the Immun-Star HRP Chemiluminescent Substrate Kit (Bio-Rad).

2.2.6 Yeast protein extract preparation

For whole yeast proteome digest, wild-type SILAC yeast strain (BY4741 lys1Δ::KAN, lys2Δ::KAN, arg4Δ::KAN) was inoculated into SD (6.7 mg/mL yeast nitrogen base without amino acids and $(NH_4)_2SO_4$ (Becton Dickinson), 0.68 mg/mL Triple Dropout CSM Mix -K, -R, -A (Formedium), 2% (wt/vol) glucose, 0.01% (wt/vol) Adenine, 0.2 mg/mL Pro, 0.02 mg/mL arginine, 0.03 mg/mL lysine). Pre-cultures were grown overnight from a single colony then diluted to OD600 ~0.05 and grown to OD600 ~0.8. Cells were collected by centrifugation at 3,000 x *g* for 5 min. Cells were lysed with 2 M NaOH, 1 M β-mercaptoethanol (Sigma) for 5 min on ice, then proteins were precipitated with 50% TCA on ice for 10 min. Proteins are precipitated by centrifugation at 3,000 x *g* for 5 min and washed twice with cold acetone. Pellets were dried and resuspended in 8 M urea, 0.5% RapiGest (synthesised in-house), 100 mM Tris pH 7.5. Protein content was determined by diluting samples 20x in water, 25 µl of which was mixed with 1 mL Bicinchoninic acid (Pierce), 0.08% $CuSO_4$ (Sigma) and incubated for 1 hr at 37 °C before measuring OD562. Concentrations were interpolated from a standard curve of BSA.

Protein preparation for poly-ubiquitylated substrate purification, a 600 mL yeast culture was grown as above but to OD600 of 1.2. Cells were harvested by centrifugation at 3,000 x *g* for 5 min and resuspended in an equal volume PBS pH 7.5

and droplets of cell suspension were frozen in liquid nitrogen. Cell were lysed in a pre-cooled Spex SamplePrep 6870 freezer mill (7 cycles for 2 min at 15 cps followed by 3 min cooling). Freezer milled yeast powder was resuspended on ice in cold TBS/1% Triton X-100, 2 µg/mL Pepstatin (Sigma), 10 µg/mL Leupeptin (Sigma), 0.5 mM PMSF (Roche), 4% Roche Protease Inhibitor (made at 1 tab/mL), 15 mM IAA. Protein was clarified by two centrifugations at 17,000 x *g* for 5 min followed by ultracentrifuation at 50,000 x *g* at 4 °C for 1 hr.

2.2.7 Tandem-repeated Ubiquitin-Binding Entities (TUBEs) affinity purification of poly-ubiquitylated proteins

2 mL (500 µL packed bead volume) of HaloLink Sepharose Resin (Promega) was pre-washed twice in TBS pH 7.5/0.05% Triton X-100 and incubated at 4 °C overnight with 4 mg NiNTA enriched 6HIS-HALO-TUBE protein in TBS pH7.5. Unbound TUBEs were removed with five washes in TBS/0.05% Triton X-100. Freezer-milled ultracentrifuged yeast extract was applied to TUBEs beads and incubated at 4 °C for 4 hr. Beads were then washed five times in TBS/0.1% Triton X-100. Beads were further washed 5 times in TBS without detergent and eppendorf tube changed twice during washing to remove plastic bound proteins and detergent. Proteins were eluted in 1% RapiGest/25 mM Tris 7.5 at 50 °C for 5 min and pooled with a wash with water. Elution was repeated with Laemlli buffer to check for elution efficiency.

2.2.8 Digestion of yeast proteins and TUBEs eluate

50 µg yeast protein extract, or entire TUBEs eluate, was reduced in 5 mM tris(2-carboxyethyl)phosphine (TCEP) for 1 hr, alkylated with 5 mM chloroacetamide (ClAA)(Sigma) for 30 min, and the remaining ClAA was quenched with 2 mM DTT. The sample was diluted 10-fold and digested with 1:100 trypsin (or 5 µg for TUBEs

eluate) at 37 °C overnight. Trypsin activity was blocked with 5 mM PMSF for 2 hr at 37 °C before acidification with trifluoracetic acid (TFA) (Thermo Scientific), precipitation of rapiGest by centrifugation at 17,000 x *g* for 15 min , and purification of peptides on a C18 MicroSpin column (The Nest Group) and dried.

2.2.9 *In vitro* ubiquitylation of Miro1

The *in vitro* reaction was performed by Agne Kazlauskaite (Alessi lab, MRC, Dundee University) and was carried out as published in (Kazlauskaite et al., 2014). 2 µg Parkin was phosphorylated by incubating with 1 µg of PINK1 at 30 °C for 1 hr in 25 µL of 50 mM Tris – HCl (pH 7.5), 0.1 mM EGTA, 10 mM magnesium acetate, 1% β-mercaptoethanol and 0.1 mM ATP. For Miro1 ubiquitylation, the reaction was diluted to 50 µL of 50 mM Tris – HCl pH 7.5, 0.05 mM EGTA, 10 mM $MgCl_2$, 0.5% β2-mercaptoethanol, 0.12 µM E1, 1 µM UbcH7 and 2 µg 6xHis-Sumo-Miro1, 0.05 mM Flag-ubiquitin, and 2 mM ATP, and incubated 30 °C for 60 min. There reaction was terminated with 1% rapiGest at 95 °C for 5 min and reduced in 5 mM (TCEP) at 50 °C for 30 min. Additional Tris–HCl was added to 100 mM to ensure buffering at pH 7.5 followed by cysteine alkylation in 10 mM ClAA at 20 °C in the dark for 30 min. Samples were diluted to 0.1% RapiGest and digested with 1 : 50 w/w trypsin overnight at 37 °C. Trypsin activity was blocked with 5 mM PMSF for 2 hr at 37 °C before acidification with 1% TFA and incubated at 37 °C for 1 hr before precipitating acid-cleaved RapiGest by centrifugation at 17 000 x *g* for 15 min. Peptides were purified on C18 MicroSpin columns and dried.

2.2.10 Sortase mediated biotinylation and enrichment of biotinylated peptides

2.2.10.1 Sortase mediated biotinylation and enrichment of biotinylated peptides from Yeast

50 µg dried tryptic yeast peptides were resuspended in 20 µL sortase reaction buffer (50 mM Tris pH 7.5, 150 mM NaCl, 5mM $CaCl_2$, 2 mM DTT) including 200 µM biotin-SGGSGGSGGGGVELPKTGEEHHHHHH bait peptide and equilibrated to 20 °C. 5 µM sortase (3x point mutant) was added and incubated for 45 min before stopping reaction with 1% RapiGest and heating to 95 °C for 5min. Sample was diluted 10x in 2xTBS and excess bait peptide depleted on 45 µL NiNTA agarose (Quagen) for 45 min before collecting the flow through and pooling with 2 washes of 2xTBS/0.1%RapiGest. Peptides in NiNTA flow-through were bound to 25 µL M-270 magnetic streptavidin Dynabead solution (Invitrogen) for 45 min, washed with 2xTBS/0.1% RapiGest 5 times. Peptides were tryptically released from the beads with 125 ng trypsin in TBS at 30 °C for 15 hr. Eluate was pooled with a wash of TBS/0.1% RapiGest and peptides were purified by C18 Microspin columns.

2.2.10.2 Sortase mediated biotinylation and enrichment of biotinylated peptides from polyubiquitin enriched yeast

Dried peptides derived from polyubiquitin enriched yeast proteins were processed as for yeast but with the following exceptions. Streptavidin beads were washed in TBS/0.1% Triton X-100 followed by 6 washed in 2xTBS/0.1%RapiGest. Elution from streptavidin was performed by bait hydrolysis in sortase reaction buffer with 0.005%Rg and 5 µM sortase (re-purified sortase 3x mutant).

2.2.10.3 Sortase mediated biotinylation and enrichment of biotinylated peptides from a Miro1 *in vitro* reaction

30 µL (of 50 µL reaction) dried peptides was processed as for yeast but with the following exceptions. Sortase reaction was for only 30 min and using the biotin-SGGSGGSGGGGGVELP**E**TGEEHHHHHH bait petpide. Bait was not depleted over NiNTA, instead, peptides were bound directly to 200 µL of streptavidin bead solution, and washed. Elution from streptavidin was performed by sortase reaction buffer containing 0.01% RapiGest, 2 µM sortase (re-purified sortase 3x mutant) and G dipeptide as 500 µM. Eluate was pooled with an additional 500 mM NaCl wash.

2.2.11 Mass Spectrometry Data acquisition

Peptides were injected onto a self packed 75 µm inner diameter PicoTip Emitter (New Objective), packed with Magic C18AQ 3 µm 200 Å beads (Michrom Bioresources) which was heated to 45 °C. A Proxeon EASY-nLC or a Dionex Ulitimate 3000 HPLC delivered a 250 nL/min gradient using buffers A (0.1% formic acid, 2% acetonitrile) and B (0.1% formic acid, 90% acetonitrile) for peptide analyses. Complex samples also included the addition of 3% DMSO. Gradients were run from 1-40% B over 25 min (synthetic peptide reactions), 60 min (*in vitro* ubiquitylations) or 90 min (complex yeast samples) followed by a high solvent wash and equilibration. Data were acquired on a Velos Orbitrap mass spectrometer (Thermo Fisher Scientific), at 60k Orbitrap resolution, and with a preview scan triggering data dependent acquisition (DDA) of the top 15 precursors above a 500 precursor intensity threshold, or 30k resolution top 1 MSn for rapid peptide analyses. Peptides were isolated within a 2 m/z window for fragmentation by rapid scan ion trap CID.

2.2.12 Data analysis

LCMS raw data files were converted to mzXML format using ReAdW.exe. For assays on synthetic peptides, in-house developed software was used to extract chromatographic peak areas for manually validated monoisotopic precursor m/z, ppm tolerance, and retention time ranges. MS2 spectra were assigned to peptide identifications by searching against a *S. cerevisiae* protein database (version 2011-02-03 from SGD http://www.yeastgenome.org/) or a composite database containing protein reagent sequences and common contaminants appended to an *E. coli* database. Search engine used was either Mascot (Matrix Science) or Comet (Eng et al., 2013). Search parameters included +57.02146 Da for carbamidomethylcysteine and variable modifications+15.994915 Da for methionine oxidation, +114.0429 Da for GG, and +383.2281 Da for LRGG. Where a tryptic threonine remnant was expect, +215.0906 Da for TGG and 101.0477 Da for N-terminal theonine was also permitted as a variable modification. Peptide precursor tolerance was set to 25 ppm (or 10 ppm for Mascot) and 0.4 Da for MS2 spectra, and semi-tryptic peptides were permitted up to 3 missed cleavages. Mascot search results were filtered at a peptide score of 25. Comet search results were validated using the Trans-Proteomic Pipeline (TPP) v4.6 with peptide and protein assignments being validated at 1% FDR using PeptideProphet and ProteinProphet.

2.3 Results and Discussion

2.3.1 Conceptual design

In contrast to previous biotechnological applications, the conjugation event in this strategy aims to specifically target the polyglycine motif that exists on ubiquitin isopeptides after tryptic digest. In this strategy, which is outlined in Figure 2.4, the LPXTG sorting signal motif is embedded within a synthetic biotinylated peptide which is used as bait to fish ubiquitin isopeptides. In addition to the N-terminal biotin and sorting signal sequence, the bait peptide also possesses an arginine in the non-specific X position to enable the isopeptides to be released from streptavidin using trypsin. The consequence of proteolytic cleavage within the motif is that a threonine residue remains attached resulting in a TGG isopeptide remnant. The TGG remnant offers a useful marker to distinguish between affinity tagged and native isopeptides without interference from the bait peptide, which is a valuable utility for methods development.

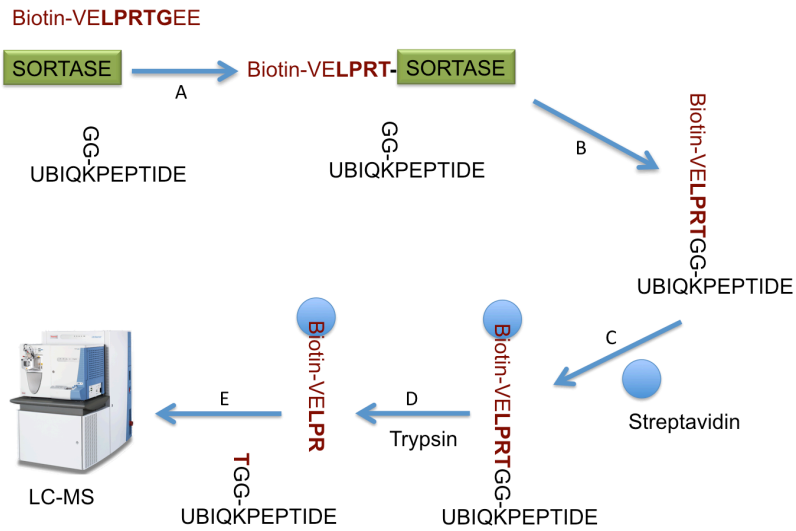


Figure 2.4 Strategy overview for the Sortase Mediated Biotinylation for Isopeptide Enrichment

Sortase two-step reaction includes proteolytic cleavage of the LPXTG motif and formation of a threonyl-thioester intermediate (A) and transpeptidation of the threonly intermediate to the diglycine nucleophile (B). Conjugated isopeptides are then enriched on streptavidin (C), then tryptically released (D) followed by LCMS analysis (E).

2.3.2 Proof of principal

Although there is evidence that sortase can conjugate to polyglycine stretches shorter than its native pentaglycine (Huang et al., 2003), there is no evidence to indicate that a branched diglycine isopeptide is a suitable substrate. To demonstrate that viability of the proposed strategy, sortase was incubated with the biotinylated bait peptide, biotin-VELPRTGEE, and a synthetic isopeptide, LIFAGK(GG)GLEDGR. Figure 2.5 demonstrates the formation of a peptide corresponding to the expected mass of the conjugated peptide. This confirmed the feasibility of biotinylating isopeptides using sortase, however, a poor conversion efficiency resulted under the initial conditions used.

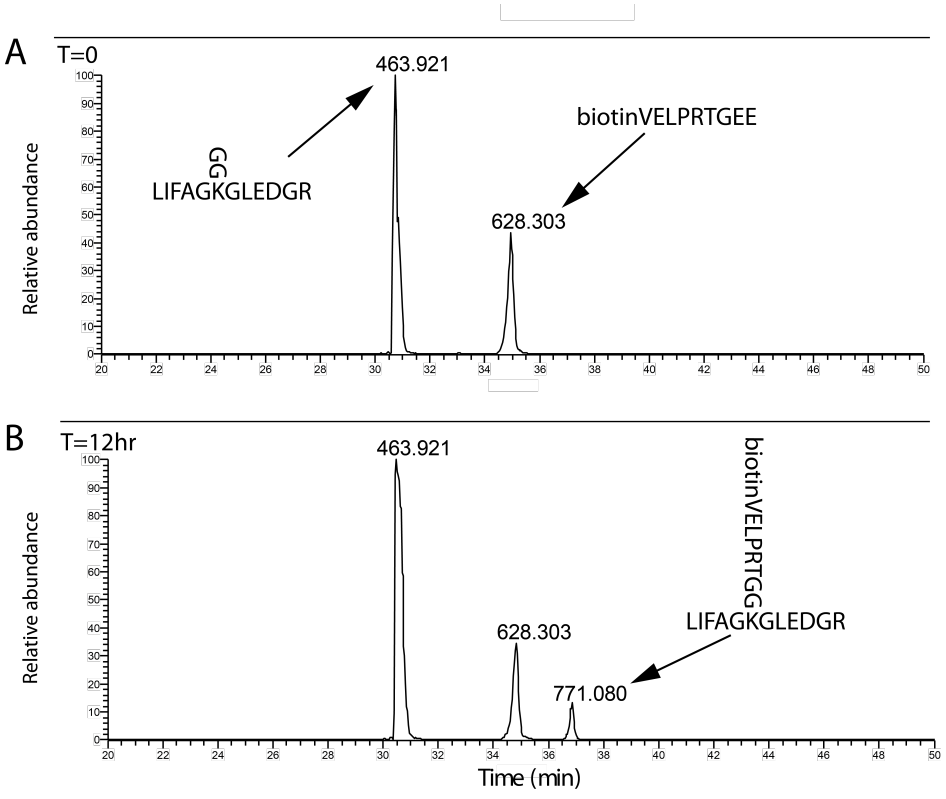


Figure 2.5 The diglycine remnant of a tryptic ubiquitin branched isopeptide is a valid sortase substrate *in vitro*

40 μM biotinVELPRTGEE (628.303 m/z, 2+), and 30 μM LIFAGK(GG)GLEDGR (463.921 m/z, 3+) isopeptide were incubated with samplings for LCMS analysis at 0 (A) and 12 (B) hours. Extracted ion chromatograms reveal the formation of the anticipated peptide conjugate, LIFAGK(biotinVELPRTGG)GLEDGR (771.080 m/z, 3+), after a 12hr incubation.

2.3.3 An excess of bait peptide is required for efficient isopeptide biotinylation

For the isopeptide enrichment strategy to be successful, a much higher conversion efficiency is required. It was noticed that the conjugated isopeptide product also contained the LPXTG motif. It is probable that this product is also a valid substrate for sortase, therefore the conjugation reaction is an equilibrium between the thioester bait intermediate and conjugated isopeptide (Figure 2.6). It was reasoned that increasing the bait peptide might promote formation of the thioester intermediate, as this is a rate limiting step (Race et al., 2009). In turn, an increased amount of thioester intermediate would drive the transpeptidation reaction in favour of biotinylating the isopeptide diglycine side chain.

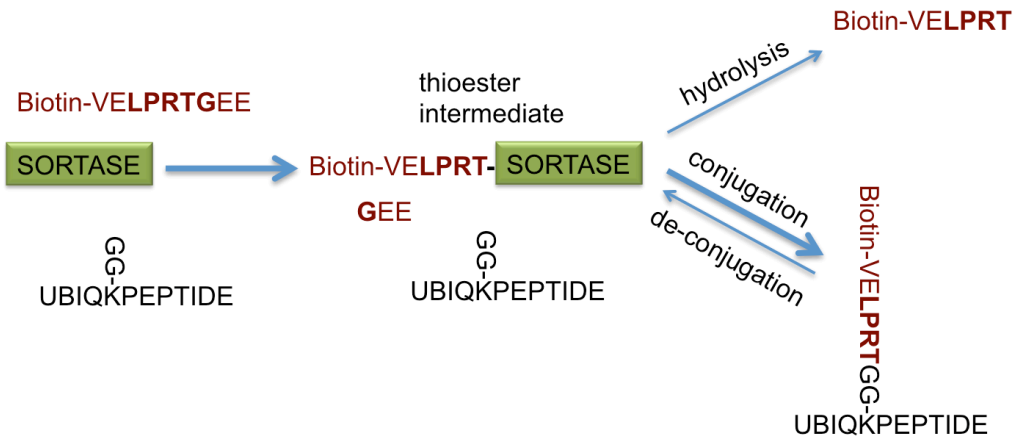


Figure 2.6 Sortase mediated isopeptide biotinylation is an equilibrium reaction. In this overview of the sortase 2-step reaction, note that formation of the thioester intermediate occurs by recognition and cleavage of the LPXTG motif present in both the bait peptide and the biotinylated isopeptide product. The bait peptide thioester hydrolysis is non-reversible.

To assess the effect of increasing bait concentration, the sortase reaction was assayed from 0-400 µM bait with a static 50 µM isopeptide. At high concentrations of bait peptide, a hydrolysed form of the bait peptide became evident, which partially co-eluted with the intact bait peptide (Figure 2.7A). This species no longer contains a

recognisable LPXTG motif and will therefore be a non-reversible product and will accumulate as the reaction proceeds. Each sample was analysed by LCMS to monitor peptide peak areas of peptide substrates and products (Figure 2.7B). Increasing the bait peptide concentration resulted in more efficient isopeptide biotinylation as indicated by both increasing biotinylated isopeptide product formation and decreasing unconjugated isopeptide concentration. The substrate formation curve plateau suggests that the system is approaching substrate saturation. Addition of even higher bait concentrations would not result in a significant increase in isopeptide biotinylation. An increase of the hydrolysed bait also occurs with increasing bait peptide, which appears to be a non-linear increase. A hydrolysis rate increase would be expected as the relative availability of diglycine substrate decreases. If a nucleophilic substrate is unavailable, a prolonged unstable thioester intermediate would result in increase hydrolysis from water.

It is clear that a many-fold excess of bait is necessary to achieve an optimal bait conversion. A large excess of the biotinylated bait peptide will also increase the scale of the downstream purification. A balance between reaction efficiency and excessive biotin reagents must be met. Under the conditions used, a 4-fold excess of bait over isopeptide was considered a suitable compromise, which corresponds to a conversion of approximately 50% (as judged by the reduced isopeptide signal in Figure 2.7).

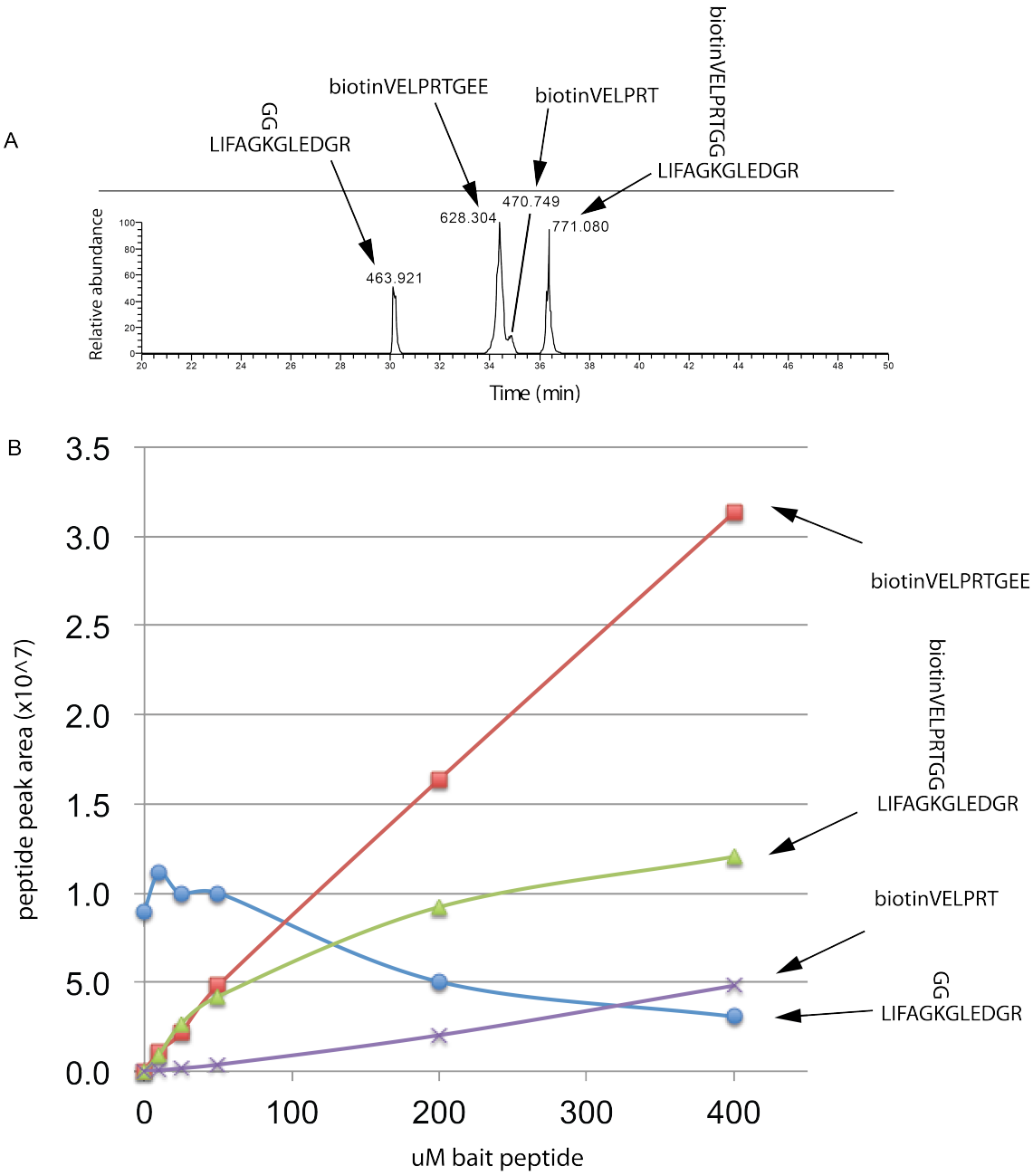


Figure 2.7 Excess bait peptide drives formation of biotinylated isopeptide

Sortase reactions were conducted for 20 hours against synthetic isopeptide with increasing amounts of bait peptide. An LCMS chromatogram of with a high bait concentration (200 μM) indicates the presence of the hydrolysed bait peptide (A). Peptide peak areas for reaction substrates and products are plotted with increasing bait peptide concentration (B). Note that the GEE C-terminal end of the cleaved bait peptide is not detectable by reverse phase chromatography due to its hydrophilicity.

2.3.4 Purification of biotinylated isopeptide

The biotinylated bait peptide allows for isopeptide recovery on streptavidin beads.

Note that the total biotin in the reaction is determined by the amount of bait peptide

used, therefore the amount of streptavidin required is not determined by the isopeptide biotinylation efficiency or bait peptide hydrolysis. Figure 2.8 presents initial attempts to purify biotinylated isopeptides on streptavidin followed by release with trypsin. To quantitatively assess the recovery of isopeptides, a tryptic digest of the reaction products was compared to the tryptic elution from streptavidin by LCMS. The eluted peptide corresponded to only 8% of the expected isopeptide signal. While this is a positive validation that tryptic elution from streptavidin is a feasible means of purifying biotinylated peptides, recovery was poor. A significant improvement in efficiency was required to make this approach a practical solution to purifying ubiquitin isopeptides.

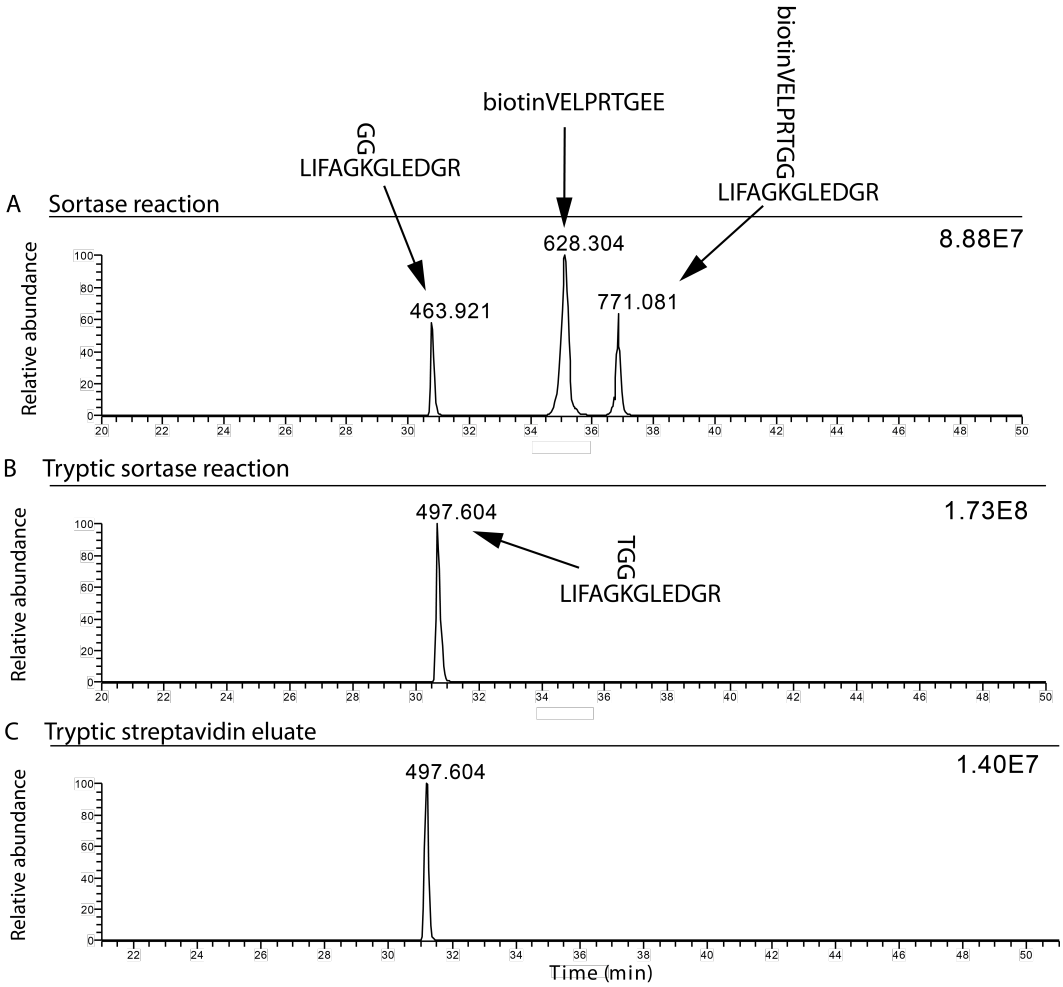


Figure 2.8 Biotinylated isopeptide can be purified and recovered with streptavidin and trypsin

Two equal portions of a sortase reaction (A) were either tryptically digested or applied to streptavidin and tryptically eluted. Comparing extracted ion chromatograms for tryptic isopeptides without streptavidin purification (B) and after streptavidin purification (C) indicated that only ~8% of the isopeptide is recovered from streptavidin. Absolute intensities are indicated to the right of each chromatogram.

The choice of straptavidin beads was also a critical factor. To maintain compatibility with LCMS analysis, beads pre-blocked in BSA or detergent had to be avoided due to a significant contamination of tryptic BSA or polymers. Complete removal of excess blocking agents from pre-blocked beads was not possible, therefore using unblocked streptavidin beads was essential. The presence of streptavidin tryptic peptides were also visible, which is an inevitable consequence of the elution procedure.

2.3.5 Optimisation of bait Peptide design

Because the recovery of isopeptides from streptavidin was a significant limiting factor, this was an initial focus for improvement. Given the short peptide sequence used, it was likely that the exposure of the tryptic cleavage site away form the streptavidin bead surface was insufficient for full access to trypsin. To improve exposure of the peptide, the bait peptide was redesigned to include a linker region between the biotin and the LPXTG motif. A selection of linkers were tested, including poly-glycine and SGG repeats which have previously been used as flexible linker between protein domains (Reddy Chichili et al., 2013). A Ttds-Linker (N-(3-{2-[2-(3-Amino-propoxy)-ethoxy]-ethoxy}-propyl)-succinamic acid) was also used as a non-amino acid alternative (Figure 2.9). Note that the arginine tryptic site was also changed to a lysine. This was to allow for testing endoprotease lysC as an eluting enzyme with dimethylation of streptavidin to protect the beads from digestion by lysC. Unfortunately, lysC was found to be inefficient at eluting isopeptides compared to trypsin.

A comparison between bait peptide designs was conducted to assess the impact on both isopeptide biotinylation and recovery from streptavidin (Figure 2.10). All bait peptides offered similar sortase reactivities. Tryptic release from streptavidin was significantly improved for all baits with extended linkers. The fact that the TtdsSGG3 linker did not result in further improvement over Ttds or SGG3 alone suggests that the shortest linker extension is sufficient to expose the tryptic site away from the bead surface. Only two thirds of the isopeptide was recovered with the longer bait peptides. The incomplete recovery is presumably non-specific losses due to additional exposure plasticware and streptavidin beads. The intermediate length SGG3 linker was selected for further use.

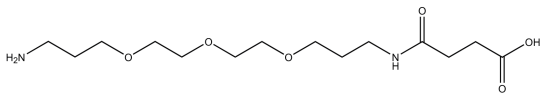


Figure 2.9 Ttds linker structure

Ttds: N-(3-{2-[2-(3-Amino-propoxy)-ethoxy]-ethoxy}-propyl)-succinamic acid, C14H28N2O6 (MW: 320.4)

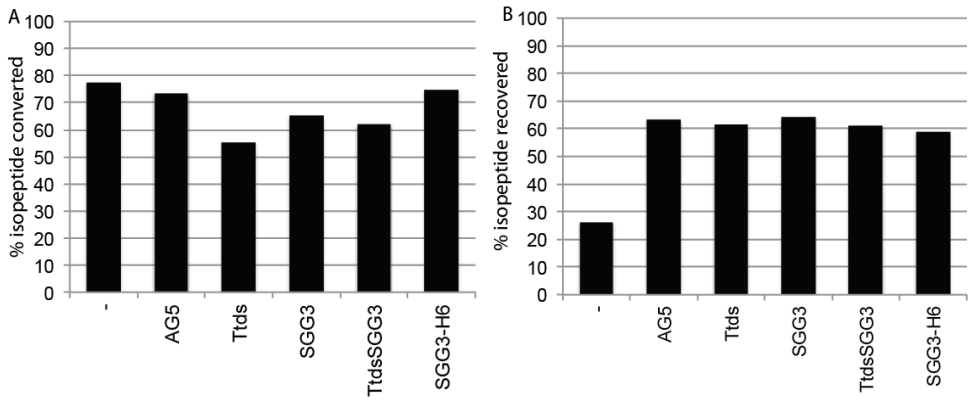


Figure 2.10 An extended bait peptide linker enables efficient isopeptide recovery from streptavidin

50 μM synthetic isopeptide was Sortase reacted for 30 hours with 200 μM of 6 different bait peptides. The conversion efficiency was determined as the TGG tryptic isopeptide signal as a percentage of total (GG + TGG) isopeptide signal in a tryptic digest of the reaction (A). The final TGG tryptic elution product was quantified as the percentage of the TGG signal in a digested portion of the sortase reaction (B).

Bait peptide key:

-	Biotin-VELPKTGEE
AG5	Biotin-AGGGGG-VELPKTGEE
Ttds	Biotin-Ttds-VELPKTGEE
SGG3	Biotin-SGGSGGSGGGGGVELPKTGEE
TtdsSGG3	Biotin-Ttds-SGGSGGSGGGGGVELPKTGEE
SGG3-H6	Biotin-SGGSGGSGGGGGVELPKTGEEHHHHHH

2.3.6 Depletion of excess bait with a C-terminal HIS6 affinity tag

Recall that efficient biotinylation of isopeptides requires an excess of bait peptide to drive the equilibrium. This excess also increases the amount of streptavidin beads required during the downstream affinity purification step. Streptavidin has a low level trypsin sensitivity and an excessive use of streptavidin results in dominating streptavidin peptides. Also included in Figure 2.10 was the SGG3-H6 bait peptide containing a C-terminal hexahistidine affinity tag. The hexahistidine modified bait offers a way of depleting unreacted bait peptide. Note than the affinity tag is released during formation of the bait-sortase thioester intermediate but remains attached to any unreacted bait. The excess GEEHHHHHH C-terminally cleaved peptide does not interfere with LCMS analyses because, like the former GEE peptide, it is hydrophilic and washes off C18 at very low solvent concentration. The bait hydrolysis will also cleave off the hexahistidine tag but without conjugating to an isopeptide, so depletion of all non-conjugated bait is not possible. Although this depletion step is optional, it permits increased amounts of bait peptide to be used, or scaling up a purification, while minimising the amount of streptavidin required in the final purification step.

Excess bait can be depleted on NiNTA, and only the flow-through is required for further processing therefore NiNTA elution can be omitted. NiNTA agarose was determined to have the capacity of approximately 2.5 μg bait peptide peptide/100 μL NiNTA agarose bead solution (Qiagen). Excess bait depletion was also explored by C18 reverse phase cutoff. C18 offers an attractive option due to its fractionation efficiency and low non-specific sample losses. This option became available only due to the hydrophilicity of the hexahistidine tag which was predicted to have very low affinity for C18 (Krokhin and Spicer, 2009). Note that in Figure 2.11A, the intact bait

peptide is an early eluting species, in contrast to the bait without a hexahistidine tag (Figure 2.8A) but still eluted later than predicted. The elution of the bait was effective at 12% acetonitrile, whereas isopeptide and conjugated isopeptide were retained (Figure 2.11B-D). Biotinylated isopeptides will have an increased retention time preventing loss of most peptides, although very hydrophilic isopeptides will still likely be lost during this procedure.

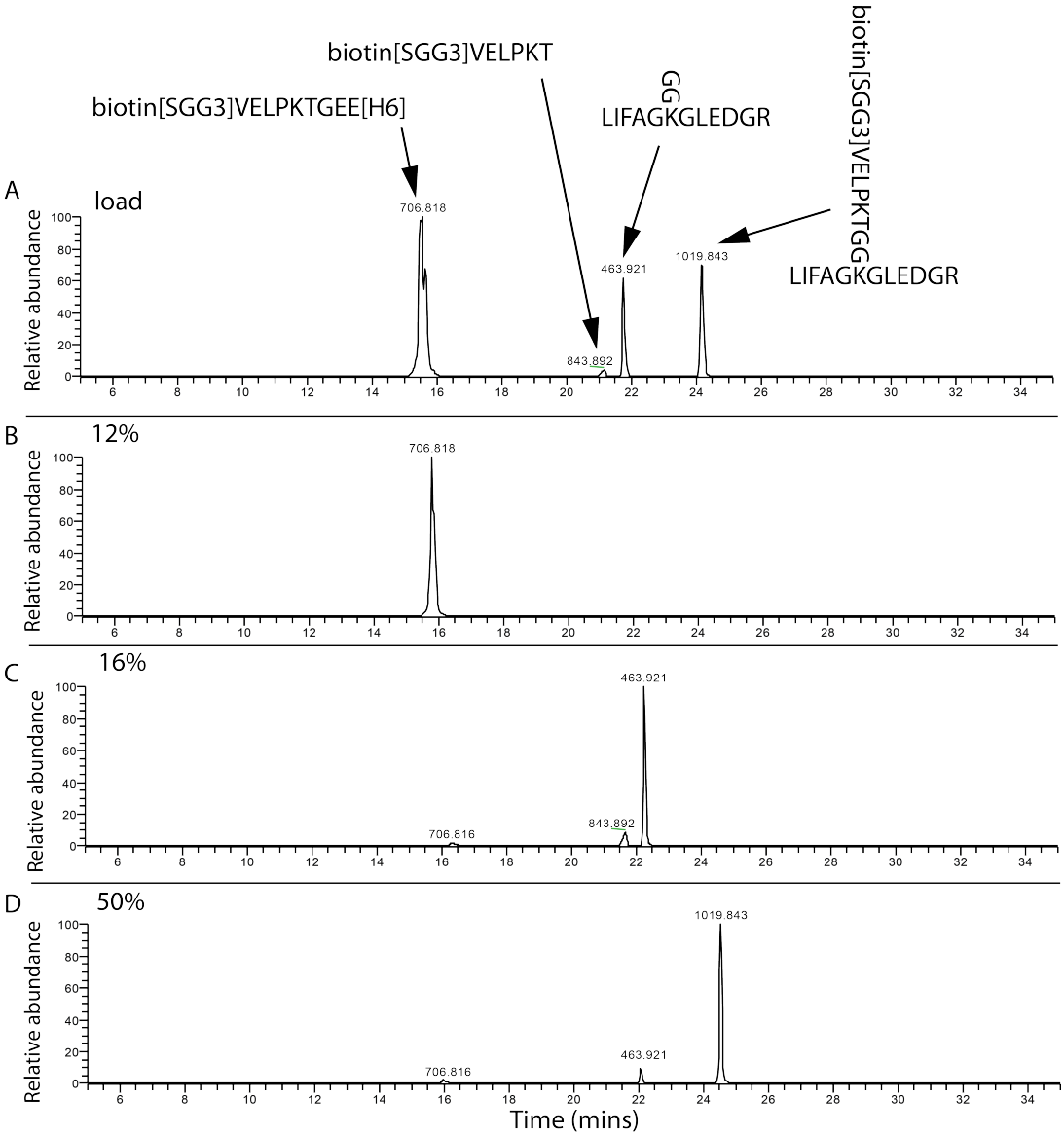


Figure 2.11 Excess bait peptide can be depleted by C18 fractionation

A sequential elution of a sortase isopeptide reaction (A) from C18 indicates that the unconjugated hexahistidine tagged bait peptide (706.82 m/z, 4+) can be depleted from C18 reverse phase media at 12% acetonitrile/0.1% formic acid (B) while retaining unconjugated (463.92 m/z, 2+) and conjugated isopeptide (1019.84 m/z, 2+), which can be eluted for further analysis (C-D).

2.3.7 Improved biotinylation efficiency using Sortase mutants with increased catalytic activity

It was observed that reactions were very slow and benefited from extended reaction times. Given that the reaction is an equilibrium, a complete isopeptide biotinylation is not achievable. However, improved isopeptide biotinylation may be achieved by ensuring that the optimal reaction time point is reached. For the current protocol, this was expected to be an impractical reaction time of more than 24 hours. Overcoming this limitation was investigated by characterising the performance of sortase reactions over time, for both wt sortase and for sortase mutants. Sortase mutants with increased bait binding affinity and catalysis were published in a screen that used sortase to exemplify a directed evolution methodology (Chen et al., 2011b). Two of the reported mutants were selected for expression and purification. A triple-point mutant (D160N/K190E/K196T) and a five-point mutant (P94R/D160N/D165A/K190E/K196T), both of which were reported to have an increased catalytic activity and decreased Km for the LPXTG motif. Assays for wt and both mutants were conducted with time points subjected to LCMS analysis to quantify product formation (Figure 2.12).

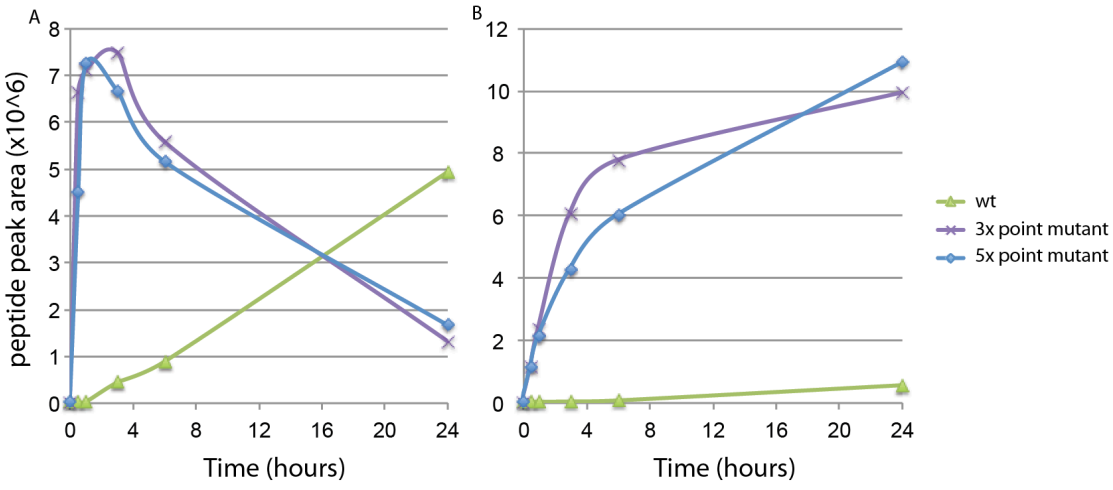


Figure 2.12 Mutant sortase enzymes enhance the rate of isopeptide biotinylation

LCMS analysis of time course reaction between 200 μM bait peptide and 50 μM isopeptide for three sortase variants follows (A) the biotinylated isopeptide product (1019.84 m/z, 2+) and (B) the hydrolysed bait peptide (843.89 m/z, 2+). Chromatographic features are indicated in Figure 2.11.

Both mutant enzyme have vastly superior reactions rates relative to wt, which is consistent with published sortase catalytic activities (Chen et al., 2011b). The wt sortase does not reach optimal product formation even after 24 hours, whereas the point mutants enabled the reaction to be completed in 1-2 hours. The expected optimal time for the wt sortase can be extrapolated to approximately 35 hours, assuming the same amount of product is produced for all enzymes. For the sortase mutant reactions, the isopeptide product decays after prolonged reaction times. This can be explained by the eventual hydrolysis of the bait peptide (Figure 2.12B), which lacks a complete LPXTG motif. As the LPXTG motif substrate availability falls below the concentration of isopeptide, the equilibrium will shift in favour of deconjugating the biotinylated isopeptide. The triple-point mutant was selected for further use in the SoMBIE method, although either sortase mutant would be suitable for accelerated substrate biotinylation.

2.3.8 Optimisation of isopeptide elution

In Figure 2.12, it was observed that the enhanced sortase enzymes will hydrolyse the bait peptide given extended reaction times. By manipulating the reaction conditions, the equilibrium can be shifted in favour of releasing isopeptides from the bait peptide. Although this is an undesirable reaction when biotinylating isopeptides, it does highlight the fact that sortase can be used as a protease with specificity for the LPXTG motif. Because the hydrolysis reaction is slow compared to the transpeptidation reaction, eluting isopeptides with a competing peptide may prove to be more efficient

than hydrolysis. A short glycyl-glycine dipeptide was tested for its ability to conjugate to the bait peptide. Figure 2.13 shows formation of an abundant glygly-conjugated bait peptide indicating that the dipeptide is an effective substrate and can be used to aid elution of isopeptide. The dipeptide is sufficiently hydrophilic to be depleted during C18 cleanup and an excess of this peptide will not interfere with downstream LCMS analysis.

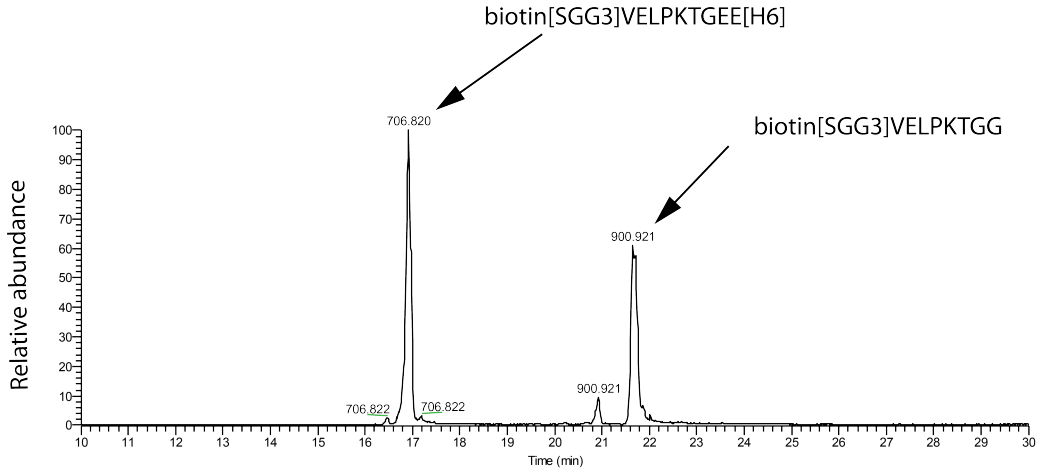


Figure 2.13. A glycyl-glycine dipeptide is a sufficient sortase substrate for competitive elution of isopeptides

Extracted ion chromatogram for reaction products after a 1 hour sortase reaction between 200 μM bait peptide and 1 mM glycyl-glycine dipeptide (not detected). An abundant diglycine-conjugated peptide confirms that a short GG dipeptide is an effective sortase substrate.

In the new elution scheme (Figure 2.14), sortase releases isopeptides by forming a thioester intermediate with the streptavidin bound bait peptide. Reforming the isopeptide-bait conjugation is discouraged by shifting the equilibrium towards both hydrolysis and conjugation to excess diglycine. The proteolytic specificity means that other protein contaminants such as keratin or proteins used to block beads from non-specific losses (eg BSA) will also remain undigested. Contaminating proteins will be depleted during peptide C18 clean-up as large proteins tend to bind very strongly to C18. Streptavidin will also remain undigested making the elution much cleaner than

with trypsin. Because streptavidin peptides will no longer dominate a sample, this makes the NiNTA or C18 bait peptide depletion step less important unless the reaction is scaled up to large volumes. As cleavage now occurs after the threonine, isopeptides can also be recovered with their original GG tryptic remnant.

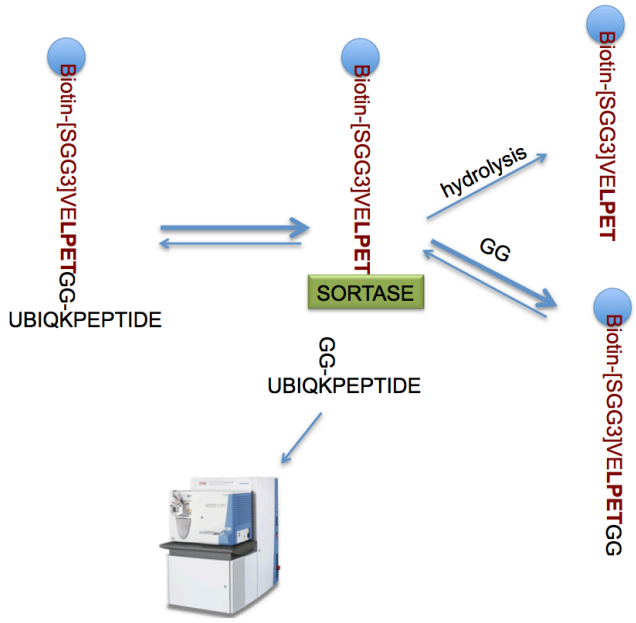


Figure 2.14 A revised strategy overview for elution of isopeptides from streptavidin using sortase as a protease.

The streptavidin bound biotinylated isopeptide (left) can be eluted with sortase. To discourage the reverse reaction, which is equivalent to the original biotinylation reaction, excess GG dipeptide is added to drive the reaction towards original product back to deconjugated form.

Now that trypsin is no longer required for isopeptide elution, the tryptic site in the non-specific position in the LPXTG motif is not necessary. This point is of importance as the bait peptide tryptic site is extremely sensitive to cleavage from residual tryptic activity. Although this is not a concern for synthetic peptide experiments, tryptic digests require PMSF treatment and C18 cleanup to deplete all tryptic activity. Failure to completely deplete residual tryptic activity results in cleavage of the bait prior to conjugation and cleavage of biotinylated isopeptides, which was observed as TGG-isopeptides prior to streptavidin enrichment and in streptavidin flow-through. Although the tryptic TGG-isopeptide method is useful for methods development, changing the

motif from LPKTG to LPETG for routine samples streamlines sample processing and improves robustness of the method. A new bait peptide was resynthesised without the tryptic site; biotin-SGGSGGSGGGGGVELPETGEEHHHHHH (SGG3-E-H6).

Isopeptide elution is the last step and is immediately followed by LCMS analysis. Unlike trypsin, the sortase enzyme was found to be incompatible with LCMS analysis. An LCMS analysis of the initial sortase preparation revealed significant interference from peptides and detergents (Figure 2.15). These contaminants are inconsequential for the first biotinylation step as these contaminants are washed prior to elution from biotin. To make an LCMS grade sortase preparation, additional size exclusion chromatography, NiNTA enrichment, and dialysis was required to obtain sufficient enzyme purity. NiNTA enrichment and dialysis alone was not sufficient to remove contaminating peptides.

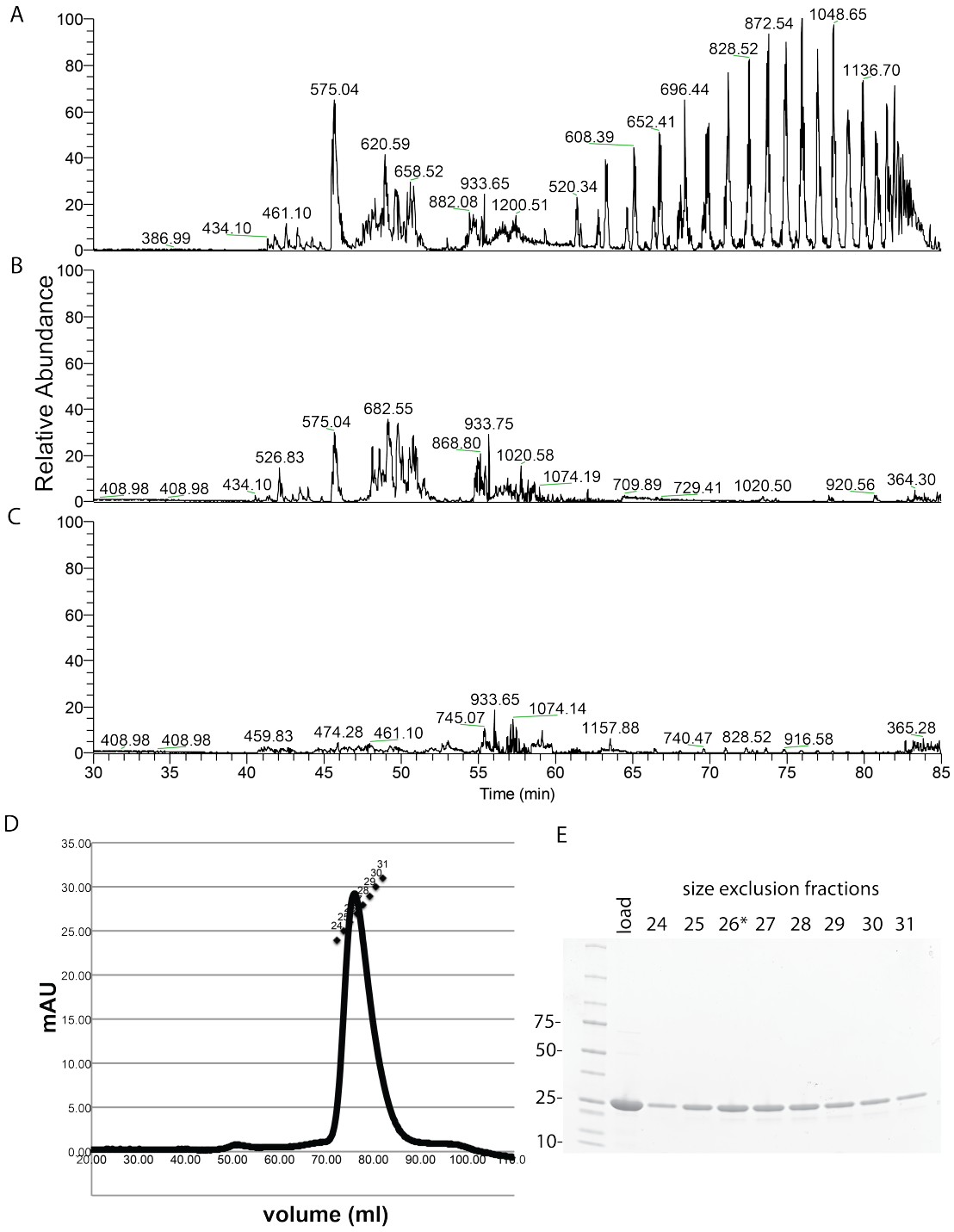


Figure 2.15 Purification of Sortase to LCMS grade quality for use as a protease

LCMS analysis of initial sortase preparations revealed contaminating detergent polymers and peptides. Re-purification was attempted with NiNTA and dialysis, and size exclusion chromatography. Base peak chromatograms of ~1 ug sortase C18 eluates (A-C) indicate initial contaminant signals (A), and effective removal of late eluting detergent polymers but insufficient depletion of early eluting peptide contaminants by NiNTA/dialysis (B). Both detergent polymers and peptide contamination were significantly reduced after size exclusion chromatography (C). Purification was conducted by selecting the most abundant fraction detected at 280nm during size exclusion chromatography (D) which was confirmed by SDS-PAGE (E, * indicated selected fraction).

2.3.9 Sensitivity improvements for low concentration substrates

Most of the previous validation experiments were conducted using 50 μM synthetic isopeptide. A more challenging sample could have a much lower isopeptide concentration due to limited sample amounts or low ubiquitylation stoichiometry. To determine conditions suitable for enrichment of low abundance substrates, time course reactions were conducted using an isopeptide concentration reduced to 1 μM (Figure 2.16). The bait peptide was monitored over two orders of magnitude corresponding to a 5-500 molar excess over isopeptide. Contrary to results observed in Figure 2.7, biotinylation occurred more rapidly and the decay of the biotinylated isopeptide was not observed after extended reaction times. A greater proportion of the isopeptide was also biotinlyated using less bait peptide than previously observed. These observations are explainable by the low isopeptide concentrations because the reverse reaction in the equilibrium is not favoured until the bait peptide concentration approaches that of the biotinylated isopeptide. The effect of reduced isopeptide concentration can be considered analogous to the excess bait peptide requirment for efficient isopeptide biotinylation (Figure 2.7) as the reaction is driven by a preferential ratio of bait to isopeptide. It is fortunate that isopeptide biotinyation becomes more efficient and less susceptible to over-reaction as the sample concentration decreases. This makes the existing protocol tolerant to reduced substrate concentration and suitable for low abundance isopeptides.

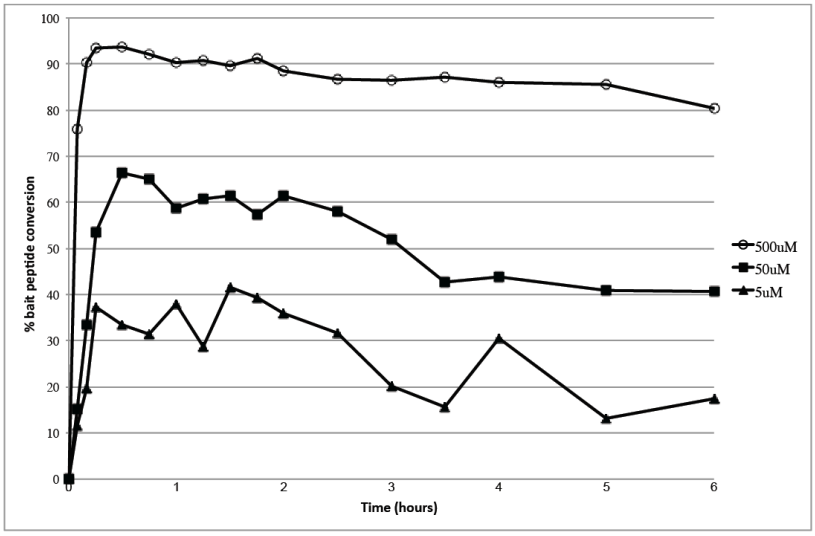


Figure 2.16 Biotinylation of isopeptide is efficient for low concentration isopeptide

A time course sortase reaction following the conversion of synthetic isopeptide (1μM) to conjugated-isopeptide using 5, 50, and 500 μM bait peptide (SGG3-E-H6). Percent conversion is measured by the loss of non-conjugated isopeptide signal and is relative to signal at time zero.

2.3.10 Application: unfractionated yeast proteome

A sortase enrichment of isopeptides from a whole yeast proteome digest was attempted. Note that this experiment pre-dates the advent of the sortase elution protocol. Biotinylation was conducted using the SGG3-H6 bait peptide, with excess bait depletion over NiNTA, and with tryptic release of peptides from streptavidin. In this experiment, a tryptic threonine remnant remains on the diglycines which permits confident differentiation between peptide carry over and enriched peptides. The most abundant isopeptides expected from the yeast proteome are the K48, K63, and K11 ubiquitin chain linkages (Peng et al., 2003). Search results for the unfractionated proteome (Mascot, peptide score >25) did not result in isopeptide identifications, which is an expected result from a highly complex non-enriched sample. In the sortase enriched sample, the reduction in sample complexity enabled confident identification of K48 and K63 isopeptides. No further isopeptides to ubiquitin or other substrates were identified. Figure 2.17 displays the pre- and post- purification LCMS analyses and extracted ion chromatograms for the identified K48 ubiquitin isopeptide.

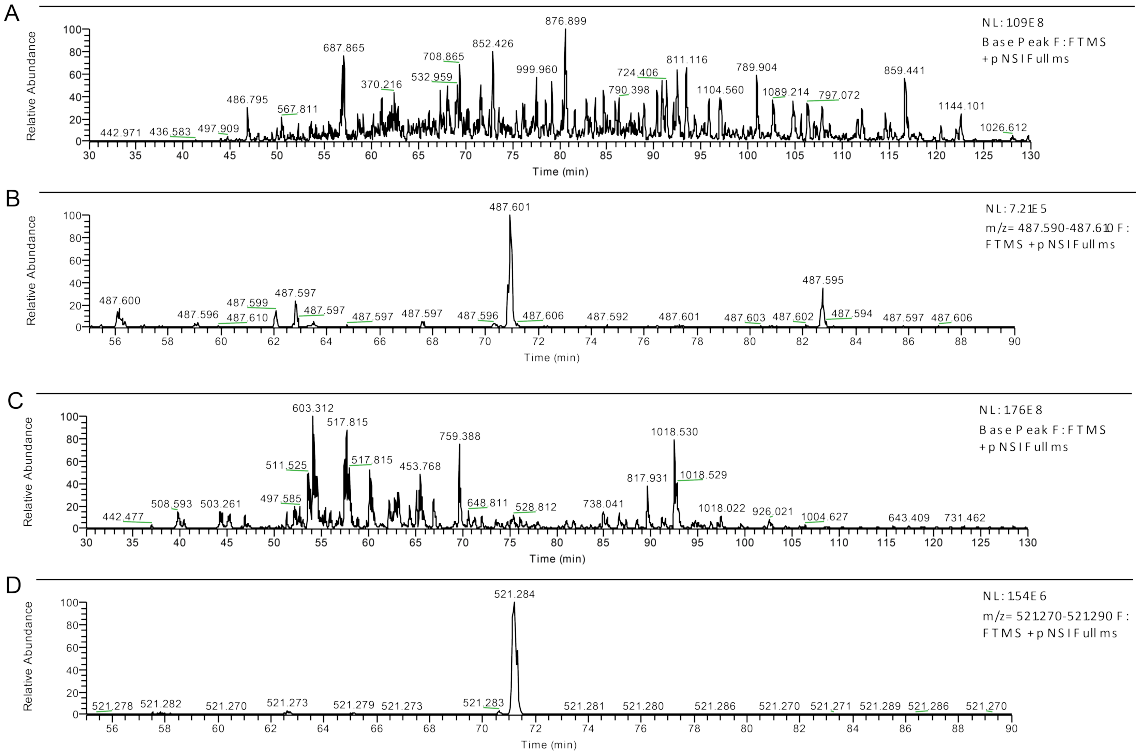


Figure 2.17 Enrichment of isopeptides from a whole yeast proteome

LCMS analyses are presented before (A-B) and after (C-D) isopeptide enrichment from unfractionated yeast. Base peak chromatogram for the enriched sample (C) indicates depletion of dominating ions observed in A. The base peak ion in C (603.312 m/z at ~55mins) belongs to digested streptavidin. Extracted ion chromatogram for the K48 ubiquitin tryptic GG-isopeptide (B) and the sortase enriched TGG-isopeptide (D) demonstrates that isopeptides are preserved during the enrichment.

One contributing factor to the poor results in this experiment is the level of streptavidin peptide contaminants. The most abundant peptide observed (Figure 2.17C) is a streptavidin peptide and this limits the amount of material that can be analysed. This highlights the importance of using sortase to elute peptides, which prevents tryptic peptides from contaminating proteins from dominating a sample. The most significant factor limiting the performance was revealed to be enrichment of peptides other than the target isopeptides. Search parameters permitted an N-terminal threonine because diglycine may also occur on peptide N-termini as well as on isopeptides. N-terminal diglycine may occur due to N-terminal ubiquitylation, most notably from linear ubiquitin chains, but also due to pre-existing KGG or RGG tryptic sites in protein

sequences. Reactivity is expected for the predicted 769 N-terminal diglycine peptides arising from 102 yeast proteins in the yeast proteome (out of 6795 predicted open reading frames). However, an overwhelming majority of peptides beginning with a single N-terminal glycine were also present indicating that there is significant reactivity towards mono-glycine peptides.

To gain a numerical overview, spectral counting was used for all peptides identified (Mascot peptide score >25). Peptides were binned into target diglycine peptides (linear and isopeptides), side reactive mono-glycine linear peptides, and non-target linear peptide without N-terminal glycine. Figure 2.18 gives a graphical overview of these peptide before and after sortase mediated enrichment of peptides. Non-target peptides are effectively depleted. Only a small number of non-target peptides remain due to carry over and also potentially tryptic peptides cleaved from enriched missed-cleaved peptides. An order of magnitude more side reactive peptides were identified than the diglycine peptides.

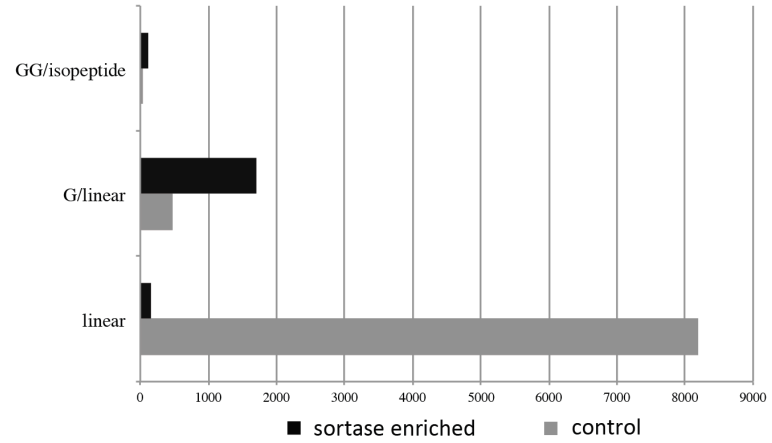


Figure 2.18 A spectral counting comparison between pre-enriched and enriched yeast proteome.

Peptide counts are compared for a pre-enrichment tryptic yeast control analysis (grey bars) and for enriched peptides (black bars). Peptide categories included sortase reactive diglycine peptides (linear and isopeptide), side-reactive N-terminal glycine linear peptides, and non-target linear peptide. Enriched peptides are identified by the threonine residue remaining after tryptic release from streptavidin beads.

These results indicate that sortase does not have the high degree of specificity towards diglycine indicated in the literature. The yeast proteome data was interrogated to determine the sequence specificity towards N-terminal amine nucleophiles. All peptides identified with an N-terminal threonine modification (indicating that the N-terminal nucleophilic amine was recognised by sortase) were subjected to a sequence analysis to determine if any positional amino acid biases exist (Figure 2.19). The results show that there is a very high degree of specificity towards an N-terminal glycine. In contrast, there is no preference for any particular amino acid in positions 2 or beyond, not even towards glycine.

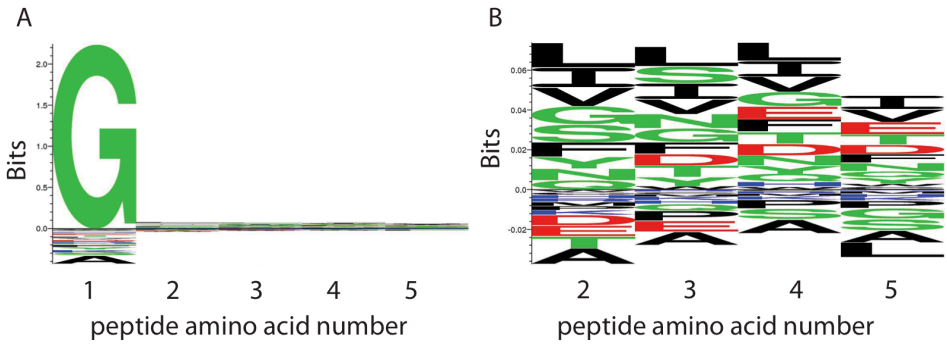


Figure 2.19 A single N-terminal glycine is necessary and sufficient for sortase recognition.

Sequence conservation analysis of purified linear peptides (n=1687) from the yeast proteome using Seq2Logo (www.cbs.dtu.dk/biotools/Seq2Logo/ (Thomsen and Nielsen, 2012)) indicates that sortase can modify peptides with a single glycine at the N-terminus (A) but has no bias towards any amino acid in subsequent amino acid positions (B).

Given the unexpected breadth of reactivity, conjugation to other amines was investigated. Analysis of *in vitro* reactions did not indicate reactivity towards Tris buffer (expected modification of +103.0633 on the bait peptide). A re-analyses of the data permitting threonine modifications on the lysine side chain revealed 118 peptides with lysine reactivity. This corresponds to ~6.5% lysine reactivity relative to single N-terminal glycine peptides. This fits well with theoretical nucleophilicity of the amines

based on the pKa of glycine (9.58) and lysine (10.67), as only the deprotonated amine is nucleophilic. Lysine reactivity could be prevented by blocking lysines prior to digesting sample, for example by dimethylation or acetylation. However, the number of peptides enriched via lysine reactivity was less than diglycine peptides, and given the extent of interfering single glycine peptides, lysine reactivity is not a limiting factor in this experiment. It is however an interesting insight into the activity of sortase. It would seem that an unbranched primary amine molecule is a sufficient nucleophile in this experimental context with peptides. The observed lysine reactivity is contradictory to previously published findings (Ton-That et al., 1998). Lysine reactivity has been reported for the sortase A orthologue in *Corynebacterium diphtheriae* during pilin assembly but onto a conserved lysine containing motif, WxxxVxVYPK (Ton-That and Schneewind, 2003). In the native context for sortase, substrate presentation or accessibility on the bacterial cell wall may play a more significant role in sortase specificity than the nucleophilic substrate preference.

2.3.11 Application: Poly-ubiquitin affinity purified yeast proteome using Tandem-repeated Ubiquitin-Binding Entities (TUBEs)

Although the SoMBIE methodology suffers from interfering side reactions in highly complex samples, potential applications may still include proteome samples pre-enriched for ubiquitylated substrates. To asses this potential, a yeast extract was enriched for poly-ubiquitin using Tandem Ubiquitin-Binding Entities (TUBEs). 6HIS-Halo-TUBEs were coupled to HaloLink Sepharose Resin before affinity purification. Western blotting for ubiquitin indicates a successful enrichment for ubiquitin and ubiquitylated substrates (Figure 2.20). Tryptic peptides from the affinity purified poly-ubiquitylated substrates were biotinylated with the SGG3-H6 bait peptide, excess bait was depleted over NiNTA, and streptavidin purified. This experiment differs from the

prior unfractionated yeast experiment in that peptides were release from streptavidin by bait hydrolysis using the re-purified sortase. This elution method will result in restoration of the original peptides and isopeptides without a threonine modification.

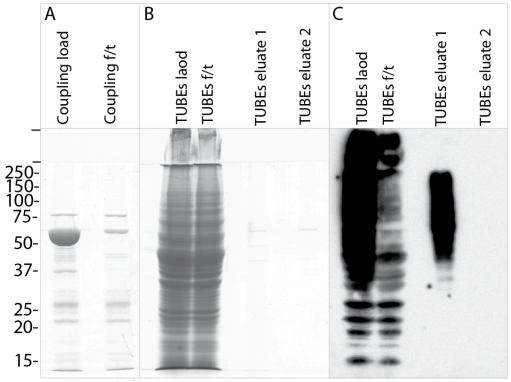


Figure 2.20 Affinity purification of yeast poly-ubiquitylated substrates

Poly-ubiquitin was affinity purified from freezer-milled yeast extract using Tandem Ubiquitin-Binding Entities (TUBEs). NiNTA purified 6His-Halo-TUBEs was coupled to chloroalkane beads. Efficient coupling is indicated by depletion from solution (A). By coomassie stained SDS-PAGE (B), a smear of protein is detected after elution in RapiGest. A western blot (FK2 anti-ubiquitin) indicates polyubiquitylated substrate enrichment and complete elution in the first eluate (C).

LCMS analyses of the TUBEs affinity purified yeast confirmed a significant enrichment for poly-ubiquitin by the identification of ubiquitin chain isopeptide linkages K6, K11, K27, K29, K48 and K63. A total of 63 diglycine isopeptides were identified over 1153 putative ubiquitin substrates (at 1% FDR). In the sortase mediated peptide purification, a similar coverage of isopeptides was observed, with a total of 67 isopeptides over 741 proteins. A spectral count analysis of peptide categories (Figure 2.21) is remarkably similar to the equivalent analysis for unfractionated yeast (Figure 2.18). Although the samples differs from the unfractionated yeast proteome, the poly-ubiquitin enriched sample is still of very high complexity, as indicated by the total proteins and total spectra identified. The spectral counting analysis indicate that the methodology still suffers from dominating mono-glycine peptides even in samples pre-enriched for poly-ubiquitylated substrates.

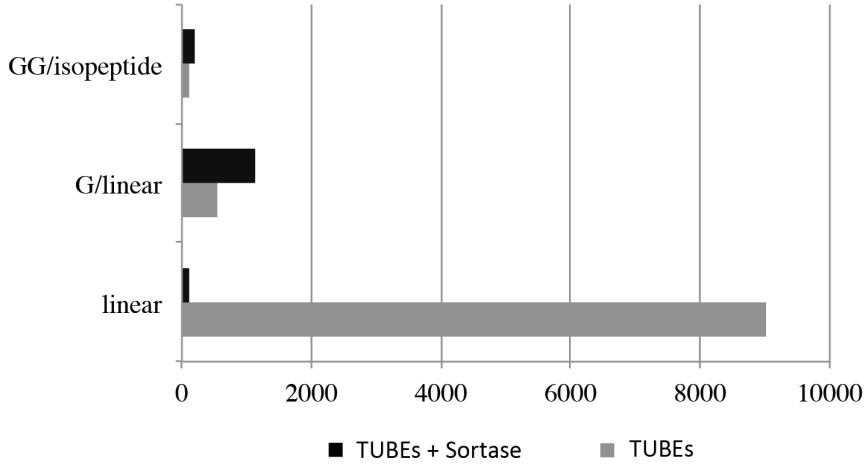


Figure 2.21 A spectral counting comparison between TUBEs poly-ubiquitin enriched and double purified TUBEs/sortase yeast proteome.

Peptide counts are compared for an LCMS analysis of poly-ubiquitin enriched yeast using TUBEs (grey bars) and for an additional peptide-level sortase purification of isopeptides (black bars). Peptide categories included sortase reactive diglycine peptides (linear and isopeptide), side-reactive N-terminal glycine linear peptides, and non-target linear peptide.

An analysis of the LCMS chromatograms (Figure 2.22) does reveal a more promising improvement in the method though. The streptavidin eluate is no longer dominated by streptavidin peptides as a result of the improved elution protocol. The enriched K48 ubiquitin isopeptide is now a major signal which indicates that the method can improve the relative isopeptide signal. Note that the base beak intensity of the sortase enriched sample was lower than in the pre-sortase sample. An increase in sample load may improve the signal to noise of low abundance peptides where substrate isopeptides would be detected. However, this would not improve the sample complexity which is currently the major limitation of the method. The SoMBIE methodology would therefore be better suited to improving the signal of isopeptides in samples of lower complexity.

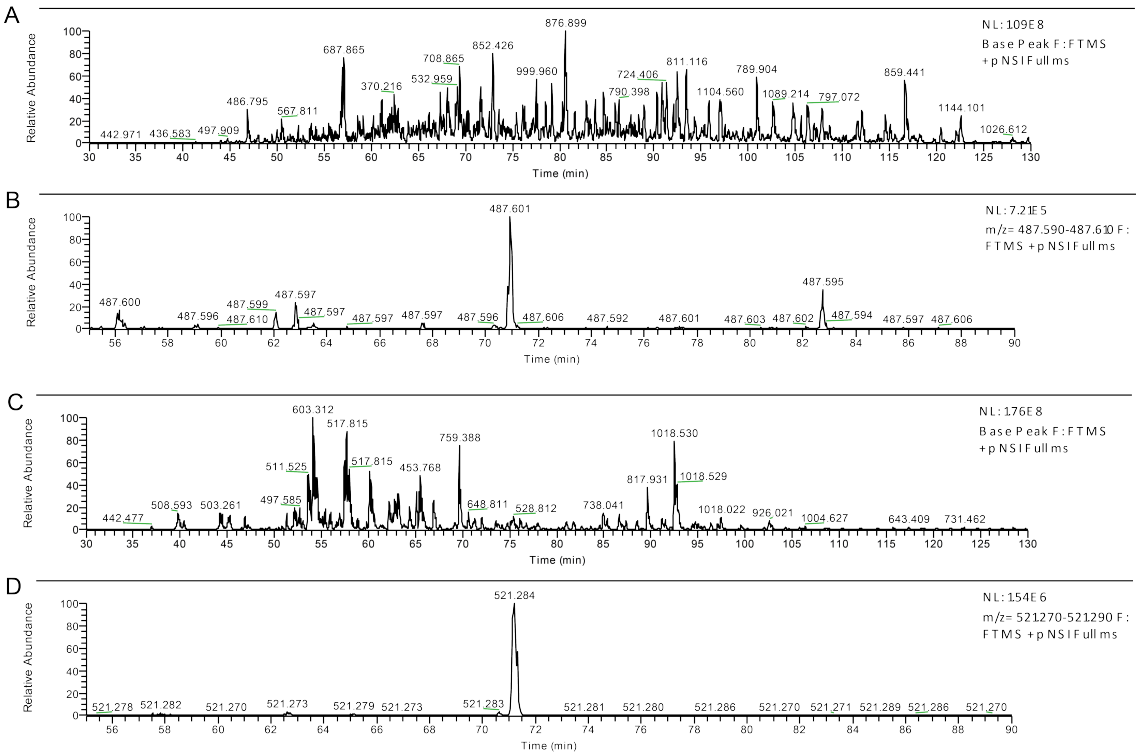


Figure 2.22 Enrichment of isopeptides from TUBEs poly-ubiquitin enriched yeast proteome

LCMS analyses are presented for a TUBEs affinity purification of poly-ubiquitin from yeast before (A-B) and after (C-D) an additional sortase mediated isopeptide enrichment. Base peak chromatogram for the enriched sample (C) indicates depletion of dominating ions observed in A. Extracted ion chromatograms for the K48 ubiquitin tryptic GG-isopeptide (B) indicates an enrichment of ~20x relative to unfractionated yeast (Figure 2.17B), and that the K48 isopeptide is a dominant signal in the sortase enriched sample (D). The other dominant peak in C is the hydrolysed bait peptide (843.894 m/z).

2.3.12 Application: *in vitro* ubiquitylation of Miro1 by PINK-activated Parkin

Mutations in both PINK1 (PTEN-induced kinase 1) and Parkin (RING-IBR-RING ubiquitin E3 ligase) are associated with early onset Parkinson's disease. PINK has been found to phosphorylate Parkin at serine-65 in the UBL domain which leads to activation of Parkin, which in turn mediates ubiquitylation of substrate proteins (Kondapalli et al., 2012). In collaborative work with Kazlauskaite et al. (2014), phosphorylated Parkin was used to ubiquitylate its putative substrate Miro1 (Mitochondrial Rho GTPase 1). Identification of Miro1 ubiquitylation sites proved to be very difficult for in-solution

digests and presented an excellent case study for using sortase to enrich for isopeptides. *In vitro* ubiquitylation was conducted as published (Kazlauskaite et al., 2014). Enrichment of isopeptide from this digest was conducted using non-tryptic bait peptide (biotin-[SGG3]-E-H6) and elution from streptavidin with re-purified sortase. The depletion step to remove excess bait was omitted.

The input to the isopeptide enrichment is presented by SDS-PAGE (Figure 2.23A). At the protein level, multiple Miro1 ubiquitylations are clearly visible, although it is unclear whether this is multi-mono ubiquitylation or ubiquitin chains attached to Miro1. The ubiquitylation does not occur in the absence of active PINK, in which case Parkin E3 ligase is not activated by phosphorylation. The PINK-WT sample was processed and a standard tryptic digest was used as a pre-sortase control. At the peptide level, the LCMS analysis of the *in vitro* reaction before isopeptide enrichment is dominated by ubiquitin peptides (Figure 2.23B). The first obvious difference in the enriched sample (Figure 2.23C) is the depletion of dominant peptides. Also note that the base peak intensity in each analysis is of similar base peak intensity ($1x10^8$-$2x10^8$) and are both near the maximum sensible load for an LCMS analysis. Spectral counting was used to determine a quantitative measure of the impact of the purification (Figure 2.24). Of the 6 proteins in the *in vitro* reaction, only isopeptides belonging to ubiquitin and UbcH7 were identified in the initial tryptic sample. After enrichment, isopeptides were identified for all proteins in the *in vitro* reaction and the spectral count was also increased for ubiquitin and UbcH7.

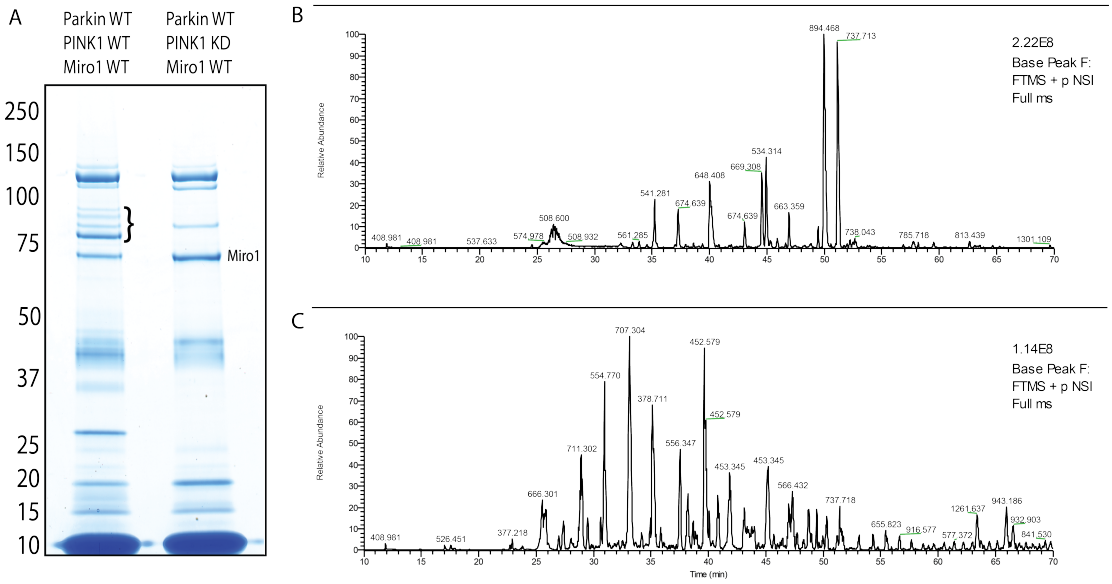


Figure 2.23 Enrichment of isopeptides from *in vitro* ubiquitylated Miro1

SDS-PAGE of Miro1 ubiquitylation reaction using WT PINK1 and kinase dead PINK1 (negative control) demonstrates multiple forms of ubiquitylated Miro1 (A). A comparison between LCMS analyses before (B) and after (C) isopeptide enrichment clearly show a depletion of the dominant peptides present in B.

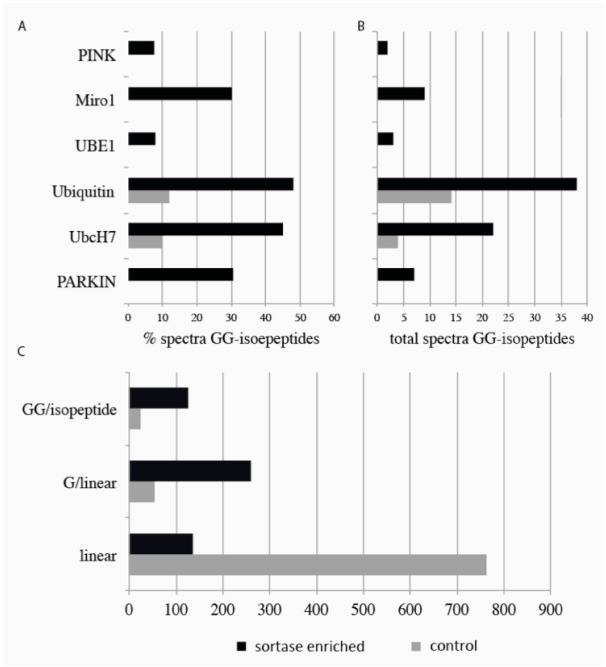


Figure 2.24 A spectral counting comparison between pre-enriched and enriched Miro1 ubiquitylation reaction.

All spectra identified at 1% FDR were counted for each protein in the sample for LCMS analyses before (grey bars) and after (black bars) isopeptide enrichment. Isopeptide spectra are presented relative to total spectra assigned to a protein (A) and as total counts (B). Spectra over the entire sample are also presented in groups of target GG/isopeptides (includes linear peptides with two N-terminal glycines), G/linear (single N-terminal glycine side-reactive peptides), and linear peptides (non-specific linear peptides without N-terminal glycine) for analyses before (grey bars) and after (black

bars) isopeptide enrichment (C). The data in C also includes E. coli peptide contaminants.

To gain a quantitative indication of the level enrichment obtained by looking at individual peptides, precursor ion chromatograms were extracted for the Miro1 lysine modifications observed. The most abundant signal was derived from a 38 amino acid peptide encompassing a modification on K572 (Figure 2.25). This lysine has been confirmed as the major ubiquitylated lysine by site mutagenesis (Kazlauskaite et al., 2014). This peptide contained a missed cleavage in addition to the expected K572 missed cleavage, making this an unusually challenging peptide to identify due to its length, high charge state and apparent low intensity. The fully tryptic cleaved peptide was not observed and is not necessarily expected to be abundant due to an aspartic acid adjacent to the lysine which is known to inhibit proteolytic cleavage by trypsin (Rodriguez et al., 2008). The signal for the identified charge state was not observed in the non-enriched sample explaining why a positive identification was not achieved. Signal for a higher charge state was detected slightly above the noise, but the higher charge state peptide did not result in an identification in either sample. The enrichment obtained for the peptide was greater than 100-fold. Additional Miro1 ubiquitylation sites were identified, including K153 and K406, and extracted ion chromatograms revealed that their precursors were not clearly detectable without enrichment (Appendix A). Additional isopeptides were assigned to Miro1 but were actually on the N-terminal SUMO1 fusion and corresponds to SUMO1 K23 and K25. There is no reason to assume that modifications to N-terminal fusion proteins have any relevance *in vivo,* which highlights the need for caution in assigning biological relevance from *in vitro* reactions.

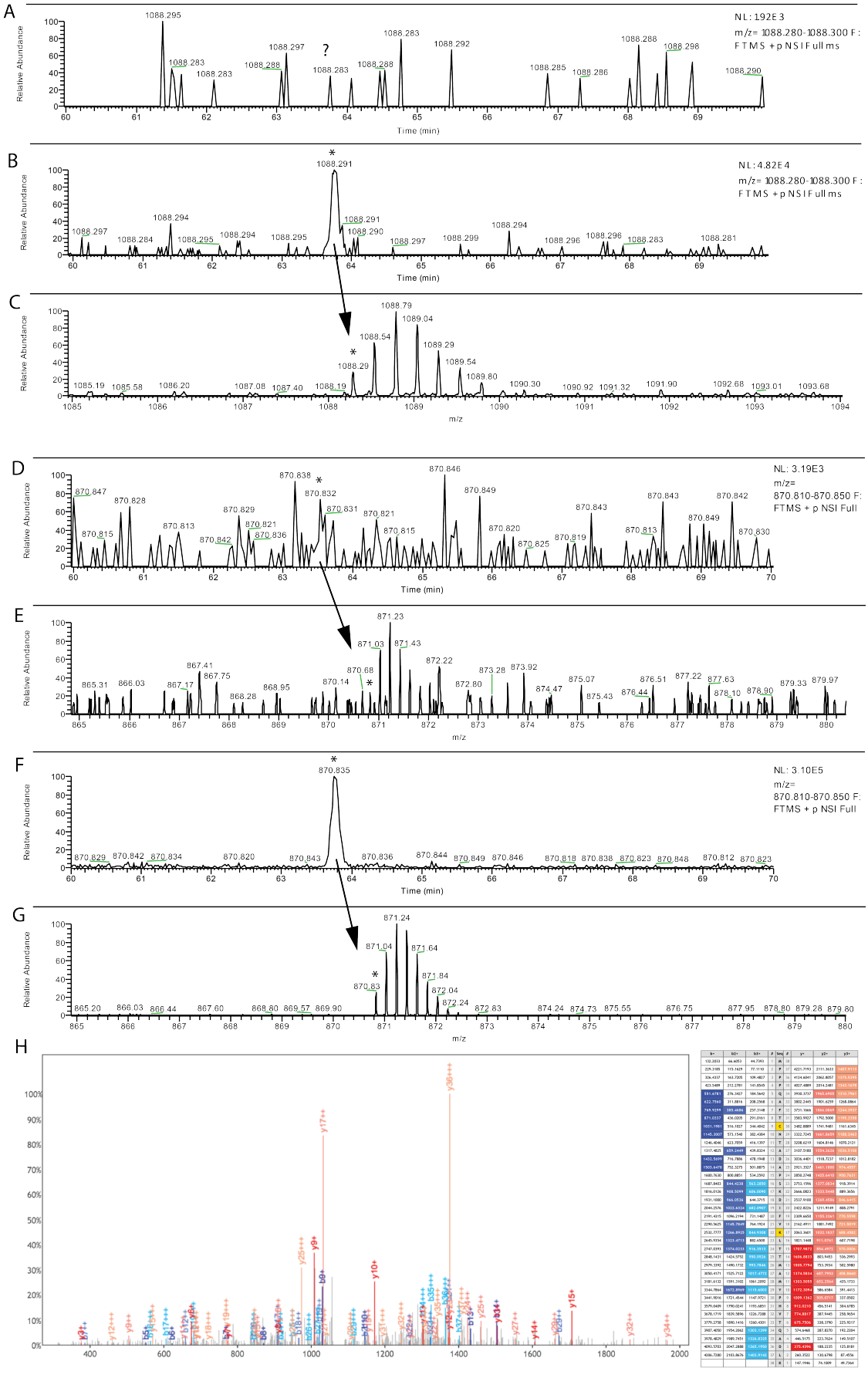

A
Relative Abundance
Time (min)
NL: 192E3
m/z= 1088.280-1088.300 F:
FTMS +p NSI Full ms
B
Relative Abundance
Time (min)
NL: 4.82E4
m/z= 1088.280-1088.300 F:
FTMS +p NSI Full ms
C
Relative Abundance
m/z
D
Relative Abundance
Time (min)
NL: 3.19E3
m/z=
870.810-870.850 F:
FTMS + p NSI Full
E
Relative Abundance
m/z
F
Relative Abundance
Time (min)
NL: 3.10E5
m/z=
870.810-870.850 F:
FTMS + p NSI Full
G
Relative Abundance
m/z
H

Figure 2.25 The SoMBIE methodology effectively enriches substrate isopeptides in *in vitro* reactions

Extracted ion chromatograms are indicated (*) for Miro1 K572 isopeptide MPPPQAFTCNTADAPSKDIFVK(GG)LTTMAMYPHVTQADLK (1088.29 m/z, 4+) before (A) and after (B) isopeptide enrichment along with the precursor spectra identified in the isopeptide enriched sample (C). Extracted ion chromatograms for the more abundant 5+ ion (870.83 m/z) was detectable before (D-E) and after (F-G) and enrichment but was not identified. The isopeptide enrichment was greater than 100x, which enabled confident identification of the K572 isopeptide (H).

The presence of linear peptides with single glycines at the N-terminus was also observed in search results for the enriched sample. In Figure 2.24C, spectral counting for this class of peptide was presented along with GG-peptides and non-target linear peptides over the entire Miro1 ubiquitylation dataset. A marked reduction of linear peptides is observed indicating a successful depletion of non-target peptides. N-terminal glycine peptides were also retained, but unlike previous experiments on complex samples, they do not heavily dominate the sample. The majority of these peptides are in fact spread over almost 100 *E. Coli* proteins which are impurities remaining from *E. coli* expressed enzyme preparations. On determining the identity of the major ions, a small number of the N-terminal glycine peptides from Miro1 and Ube1 were observed to be dominating the sample. Extracted ion chromatograms for these peptides reveal that they are not high intensity signals in the original sample, and that their dominance in the enriched sample is a result of enrichment rather than carry-over (Figure 2.26). Note that these ions correspond to dominant ions in Figure 2.23C. These peptides are also enriched approximately 100-fold indicating that the SoMBIE method results in similar levels of enrichment for ubiquitin GG-isopeptides and single glycine N-terminal containing peptides. Importantly, the cross reactivity towards single glycines did not completely counter the benefits of the procedure. This validates the use of the SoMBIE procedure for improving isopeptide detection in samples of low complexity.

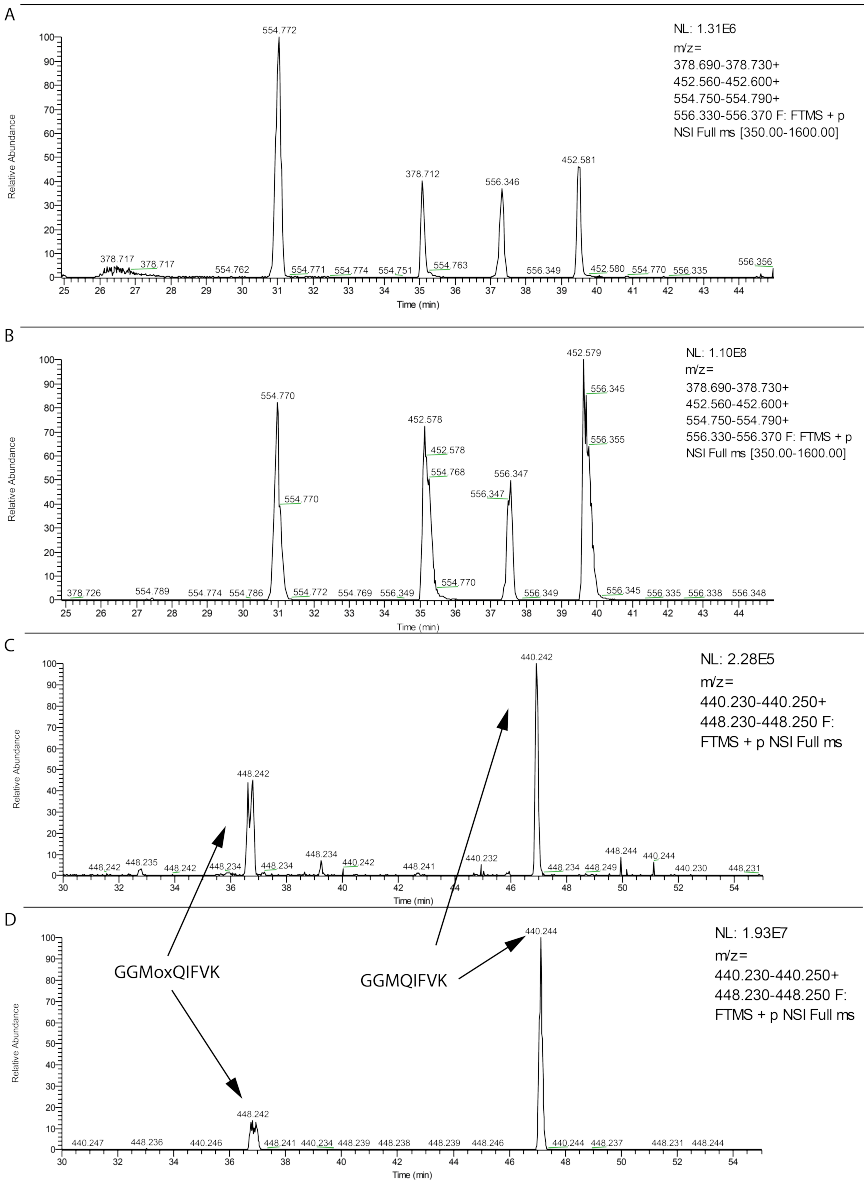


Figure 2.26 SoMBIE also enriches linear peptides with N-terminal glycine and N-terminal ubiquitin diglycine

Comparison of the four most abundant N-terminal glycine containing peptides (belonging to Miro1 and Ube1) before (A) and after (B) enrichment. Note that peptides in A are at ~1% of the signal intensity in the pre-enriched sample (refer to fig2.23A) and that the peptides in B correspond to abundant peaks in Figure 2.23B. Diglycine linear peptide corresponding to N-terminally ubiquitylated peptides are also enriched (C-D).

A class of linear peptides of importance is the N-terminal diglycine modification derived from N-terminal ubiquitylation. Purification of the N-terminal ubiquitin peptide carrying a diglycine (GGMQIFVK) was also enriched ~100x (Figure 2.26C-D). The

enrichment of N-terminal glygly is an important distinction between this method and those that are strictly specific for isopeptides. Note that this peptide exemplifies the capability of the purification but is not biologically relevant in this experimental context because it is actually an artefact of the N-terminal FLAG-tag on ubiquitin.

2.3.13 Future method improvements

Because the mono-glycine side reaction limits the methods effectiveness, improvements to this method must focus on improving the specificity of sortase towards diglycine. One approach could be to find alternative sortase variants with altered specificity. In the directed evolution strategy that discovered the sortase mutants (Chen et al., 2011b), randomly mutated sortase was expressed along with cell surface immobilised triglycine nucleophile. Biotinylated peptide containing the LPETG sorting signal motif was added to enable sortase mediated biotinylation of cells and FACS sorting was used to enrich for cells with efficient labelling. By using limiting amounts of LPETG peptide, sortase mutants with lower Km for the LPETG motif were discovered. The same procedure could potentially be repeated to obtain a diglycine specific sortase variant by expressing a cell surface immobilised LPETG peptide. The conjugation could then evolve towards recognition of biotinylated GG-isopeptide with negative screening against a biotinylated linear GX peptides. (Theile et al., 2013) points out that the sorting signal also requires an additional amino acid directly after the motif for efficient binding; the sorting signal motif is really therefore LPXTGX. The GX binding position is likely to be at or near the same binding site for the recipient nucleophile. Also, sortase is non-reactive towards a single glycine amino acid (Huang et al., 2003). Taken together, this suggests that the enzyme-substrate binding at the second amino acid residue of the nucleophilic peptide is important for substrate recognition which gives hope that the specificity of this binding may be manipulated. Also note that the evolution of sortase towards a stronger preference for a GG

nucleophile may further restrict the sorting signal motif to LPXTGG, in which case the bait peptide will need to be modified accordingly.

It is possible that the work with sortase mutants has been overly focused on the rate of reaction. A range of sortase mutants were reported by (Chen et al., 2011b) with increased catalytic rates. However, the increased Kcat and decreased Km(LPXTG) were also approximately proportional to an increased Km(GGG). (Chen et al., 2011b) reported that the mutant sortase enzymes did not have compromised specificity towards substrates of 3-5 glycines despite differences in Km(GGG). However, shorter glycine stretches were not tested and the increased Km(GGG) may translate to an increased tolerance for GX peptides. Should this be the case, very long reaction times with the wt sortase may provide improved diglycine selectivity. 100% specificity with wt sortase is not possible given reports that GA and GV peptides have some reactivity, albeit with a much higher Km values (Huang et al., 2003).

Although a strong preference for diglycine was reported (Huang et al., 2003), a lower catalytic rate for mono-glycine peptides can be overcome by having a high concentration of GX peptides, as is expected to be the case in complex proteome samples. The high degree of observed GX reactivity may be a result of reacting for a maximum yield of biotinylated isopeptide which may also be nearing a maximal reaction point for GX peptides. Re-optimising on GG/GX peptide biotinylation ratio may offer improved isopeptide recover relative to undesirable linear peptides. However, this reduced reaction time will result in reduced total recovery of isopeptides and will diminish the sensitivity of the method. This would also require a much tighter control of the reaction times and would need to revert to a slower sortase variant. Although some improvement may be achieved by taking advantage of differential

GG/GX kinetics, alternative sortase enzymes with higher specificity would be preferable.

The reactivity to single glycine peptides also highlights the reversibility of the initial thioester formation. The cleaved C-terminal portion of the bait peptide can be reconjugated to reform the initial bait peptide (refer to first step in Figure 2.6) and this equilibrium plays a part in the requirement for excess bait peptide. An alternative bait could be synthesised where the -LPXT*G- motif is made as a depsipeptide (where * is an ester bond rather than an amide). Depsipeptides are reactive in the initial protease step to form the thioester intermediate, but the resulting hydroxy-glycine peptide is not reactive in the transpeptidation step due to lacking the amine nucleophile (Williamson et al., 2012; 2014). Although this will not improve reaction specificity, it will improve the reaction efficiency thus minimising the amount of bait peptide and affinity media required. The bait peptide could also be pre-bound to streptavidin or alternatively a non-biotinylated bait could be coupled directly to NHS activated bead. Direct coupling was initially avoided as it does not allow for characterising the reaction by LCMS. Re-optimising for this approach would generate a more time efficient and simpler protocol reminiscent of a kit-based approach for ubiquitin isopeptide enrichment.

Chapter III
Isotope Coded Isopeptide Detection (ICID)

3.1 Introduction

Ubiquitin is well placed as the type specimen of the UBLs. Not only was it the first UBL discovered (Goldstein et al., 1975), it is the most well studied UBL and has the most extensive collection of know substrates. But progress in ubiquitin analysis by mass spectrometry was not achieved solely for its early discovery. Ubiquitin, as well as NEDD8 and ISG15 tryptic isopeptides retain a simple diglycine modification and are identified with relative ease. The challenge of detecting low abundance isopeptides was also well addressed with the advent of anti-diglycine antibodies. Although ubiquitin isopeptide analysis is still not as straightforward as proteomic analysis of non-modified proteins, it is in a far more advanced state than that of the other members of the ubiquitin like family.

An interest in the functional significance of the entire UBL family is also growing. In a recent study using protein arrays and application of mitotic cell extracts, more than 1500 proteins were found to be putative substrates of a range of UBLs including ubiquitin, SUMO2/3, NEDD8, FAT10, SUMO1, UFM1, and ISG15 (Merbl et al., 2013). This research revealed that each UBL has a distinct network of substrate proteins and the role of each UBL is not fully appreciated. As current research broadens its focus to the wider group of UBLs, there is an increasing requirement for mature MS-based techniques for studying each member of the ubiquitin family. One of

the major hurdles in the analysis of UBLs is the identification of isopeptides, particularly those with long tryptic remnants.

In the previous chapter, the enrichment of isopeptides was addressed in an attempt to deal with the dynamic range and sensitivity challenge of UBL analyses. In this chapter, research is presented that approaches the study of UBLs from the perspective of determining confident isopeptide identifications. A particular focus is given to UBLs with long tryptic remnants that require non-standard proteomic techniques for analysis.

3.1.1 Enrichment of UBLs.

UBL modified proteins represent a small fraction of the proteome and, like ubiquitylated substrates, require enrichment and purification to achieve detection by mass spectrometry. Of the non-ubiquitin UBLs, SUMO has received the most attention. PolySUMOylated proteins can be enriched with a SUMO interacting domain from RNF4, which contains four SUMO interacting motifs (SIMs). Although this domain binds only weakly to mono-SUMO, it has been successfully applied to identification of more than 300 putative polySUMO conjugates from HeLa cells (Bruderer et al., 2011). Mono-SUMOylated substrates can also be enriched with monoclonal antibodies raised against C-terminal peptides derived from SUMO1 and SUMO2. This approach has yielded 584 endogenous SUMO modified proteins from HeLa cells (Becker et al., 2013). In genetically tractable systems, expression of HIS6-tagged SUMO (Vertegaal et al., 2006) or TAP-tagged SUMO (Golebiowski et al., 2010) enables a purification along with its modified substrates . The high stringency purification of TAP-tagged SUMO2 has enabled detection of 766 heat shock induced changes in HeLa cells (Golebiowski et al., 2009).

In an earlier study in yeast, 251 candidates for SUMOylation were identified by mass spectrometry following a His6-FLAG-SUMO purification, six of which were confirmed through identification of SUMO isopeptides (Denison et al., 2005). Denison et al. (2005) were able to achieve these isopeptide identifications because the yeast SUMO orthologue, Smt3, has a short tryptic remnant of EQIGG. The mammalian scenario is contrasted to yeast in that mammalian SUMO has a very long tryptic remnant. In recent enrichments of SUMO in mammalian systems using polySUMO interacting domains, anti-SUMO antibody and HIS6/TAP-tagged SUMO, SUMO isopeptides were not identified. In these studies, identified proteins were presumed substrates because they co-purified with the modifying UBL. A technique for collapsing poly-SUMOylated proteins into one molecular weight even aims to remove the UBL modifier by treating with SUMO specific proteases making conjugation site identification impossible (Blomster et al., 2009). Although enrichment strategies are improving significantly, identifying conjugation sites to confirm putative substrates has largely been neglected - and with good reason as the analysis of UBLs with long tryptic remnants remains extremely challenging.

3.1.2 Spectral complexities of long UBL tryptic remnants

When dealing with most members of the ubiquitin like family, a tryptic digest does not result in an inconsequential remnant like the diglycine from ubiquitin, NEDD8, and ISG15 (Figure 3.1). A range of longer remnants occur, from a relatively short EQIGG from Smt3 (*S. cerevisiae* SUMO orgtholog) to the extreme case of human SUMO2 which gives rise to a tryptic remnant of 32 amino acids. Unmodified peptides and peptides with simple modifications are assigned identities by correlating the spectral pattern to their theoretical fragment ions. Simple modifications are accounted for simply by a mass shift in the ion series. Spectral interpretation of long UBL modified

peptides is complicated, not just because of the size of the modification, but because the modification itself is proteinaceous and prone to fragmentation (Figure 3.2). For large SUMO2 peptides, the modifier is longer than a typical tryptic peptide and resulting fragments often dominate over the substrate fragment ions.

```
Ubiquitin:  QQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG
ISG15:      LFWLTFEGKPLEDQLPLGEYGLKPLSTVFMNLRLRGG
NEDD8:      QQRLIYSGKQMNDEKTAADYKILGGSVLHLVLALRGG
FAT10:      TQIVTCNGKRLEDGKMMADYGIRKGNLLFLACYCIGG
SUMO1:      SLRFLFEGQRIADNHTPKELGMEEEDVIEVYQEQTGG
SUMO2/3:    QIRFRFDGQPINETDTPAQLEMEDEDTIDVFQQQTGG
Smt3_sc:    SLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG
Ufm1:       SAIITNDGIGINPAQTAGNVFLKHGSELRIIPRDRVG
```

Figure 3.1 C-terminal sequences from a selection of Ubiquitin like modifiers.
UBLs present a variety of C-terminal sequences resulting in a diverse range of tryptic remnants. Digestion with trypsin cleaves after arginine or lysine to leave a C-terminal portion on the lysine of a substrate tryptic peptides (underlined).

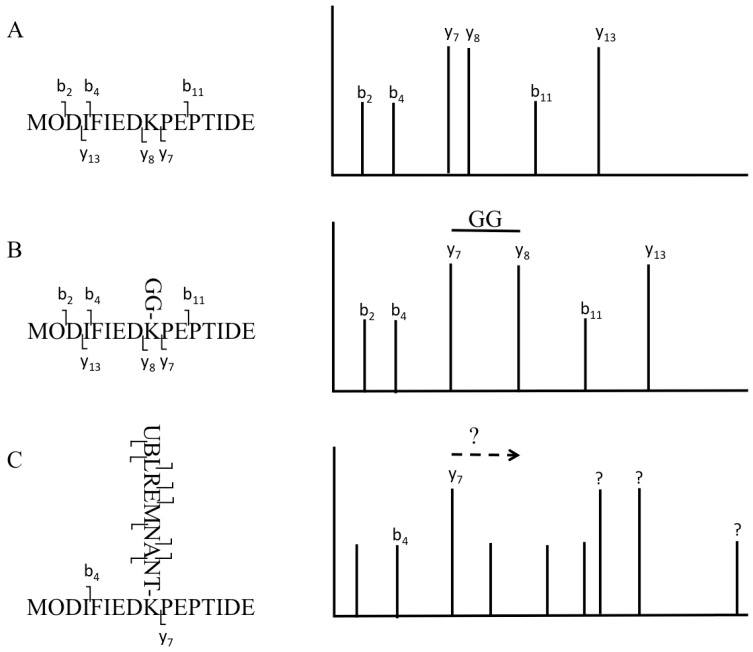


Figure 3.2 Fragmentation of peptides and modified peptides
A depiction of fragmentation for a peptide with no modification (A), a ubiquitin tryptic remnant (B), and a long UBL tryptic remnant (C). The long UBL modified peptide has altered fragmentation and is mixed with a series of b and y ions overlapping with the substrate ion series, and in many cases dominate over fragment ions from the substrate peptide.

3.1.3 Alternative proteolytic enzymes:

The use of alternative proteolytic enzymes is an useful strategy to improve protein coverage in regions where tryptic sites are rare or overly abundant. This approach is also useful for the analysis of isopeptides with long tryptic UBL remnants because the generation of short remnants are amenable to analysis using standard standard proteomic software. The combination of trypsin, Glu-C, Asp-N, and pepsin was used to improve coverage of SAE2 to identify auto-SUMOylation sites (Truong et al., 2012). Chymotrypsin has proven useful for detecting a shorter QQQTGG SUMO1 remnant (Dumont et al., 2011). Similarly, elastase produces a range of remnants including GG for SUMO1 and also QTGG and TGG for SUMO2 (Chicooree et al., 2013c). Acid hydrolysis at aspartic acid also produces a relatively short VFQQQTGG for SUMO1 (Osula et al., 2012). The effect these alternative enzymes also have on the substrate peptide should not be ignored. Although Smt3 already has a short tryptic remnant, identifying lysine conjugation sites on UBC9 still requires the use of GluC to identify isopeptides (Klug et al., 2013).

Many of the alternative enzymes, such as chymotrypsin, elastase, and pepsin, have a broader specificity than trypsin. This results in a more complex sample due to miss-cleavage of the UBL remnant or substrate peptides creating multiple peptide species. Trypsin has high specificity and maintains a C-terminal lysine/arginine charge making it particularly desirable for complex samples. The use of alternative enzymes for analysing SUMO isopeptides has almost exclusively been for relatively simple *in vitro* reactions.

3.1.4 Sequence manipulations for improved analytic capabilities:

The use of SUMO mutants is becoming a popular approach to study SUMO isopeptides. By inserting a tryptic site near the C-terminus, the length of the remnant can be shortened enabling analysis using standard proteomic search engines while maintaining a tryptic sample processing. A variety of mutants have been created, such as RGG to generate a ubiquitin-like di-glycine remnant (Knuesel et al., 2005; Wohlschlegel et al., 2006). A particularly effective strategy includes a series of mutants in each of the SUMO variants that allows each to be distinguished by their unique short tryptic remnants (Figure 3.3). Sequence modification has also been taken to a further extreme by creating a lysine-deficient SUMO2 in addition to Q87R and T90R mutations. This creates a lysC resistant SUMO resulting in digestion of all proteins except SUMO and its conjugated peptide. Subsequent SUMO enrichment and tryptic digest enables substrate peptides to be analysed with reduced sample complexity. This strategy enabled detection of 103 substrate isopeptides (Matic et al., 2010).

SUMO has also been altered for isopeptide purification purposes. By introducing cysteines into the C-terminal peptide of SUMO between lysC and tryptic sites, lysC generated isopeptides can be purified on thiopropyl beads and then release of diglycine isopeptides with trypsin (Blomster et al., 2010). This method has the benefit of both a short remnant and a means of purifying isopeptides. Although this required alteration of six amino acids in the SUMO sequence, conjugation to a substrate protein PARP-1 was shown to be viable *in vivo*. Manipulation of SUMO to generate a diglycine remnant is not just useful for shortening the isopeptide remnant for bioinformatic simplicity. The ubiquitin-like diglycine remnant has proven most successful in combination with the ubiquitin anti-diglycine antibody (Tammsalu et al., 2014). This

strategy enabled identification of more than 1000 sumoylated lysines from HEK293 cells stably expressing 6His-SUMO2-T90K.

The manipulation of SUMO sequences has significantly enhanced analytical power for cells that are genetically tractable. However, when we consider UBLs in light of human disease, we are faced with significant constraints. Tissues from human patients will contain wild-type UBL sequences and genetic manipulation and transfection of biologically relevant primary cell lines is difficult (Gresch and Altrogge, 2012). To study the entire UBL family in medically relevant scenarios, it is essential that we develop the ability to study UBLs with their long tryptic remnants intact.

```
SUMO1:  …FEGQRIADNHTPKELGMEEEDVIEVYREQTGG
SUMO2:  …FDGQPINETDTPAQLEMEDEDTIDVFRQQTGG
SUMO3:  …FDGQPINETDTPAQLEMEDEDTIDVFRNQTGG
```

Figure 3.3 C-terminal arginine mutants enable effective MS analysis of SUMO isopeptides.

Strategy for creating mutants of SUMO1, SUMO2, and SUMO3 demonstrate ability to produce short tryptic remnants and differentiate between SUMO variants (Galisson et al., 2011; Lamoliatte et al., 2013) . Mutated amino acids are bold/underlined and were formally all Q.

3.1.5 Methods for improved UBL isopeptide detection

The UBL remnant presents a constant feature across this class of peptide. As a result, Lamoliatte et al. (2013) observed that the remnants of SUMO C-terminal arginine mutants had characteristic fragment ions that could be used for identification and confirmatory purposes. By repeatedly cycling through consecutive 100-*m*/*z* precursor isolation windows, detection of diagnostic ions from the NQTGG remnant allowed for a targeted re-acquisition of SUMOylated peptides from HEK293 cells expressing the SUMO3 C-terminal arginine mutant. The detection of UBL isopeptides

through diagnostic remnant ions can also be enhanced by chemical modification. Reductive methylation of peptide N-termini enhances small a- and b-ions making even the short ubiquitin GG remnant ions detectable in MS2 spectra (Chicooree et al., 2013a; 2013b). These methods improve identification of isopeptides with short remnants, however, they have not been demonstrated for more complex isopeptides like native SUMO isopeptides.

A method has also been developed for distinguishing isopeptides from linear peptides by guanidinating lysines with O-methyl isourea and sulfonating N-termini with amine reactive 4-sulfophenyl isothiocyanate (Wang and Cotter, 2005; Wang et al., 2005). MS2 spectra display dominant ions for the loss of the tag and ubiquitin isopeptides have a characteristic loss of two tags, one from each N-termini. On recognising the isopeptide signature in MALDI spectra, spectra could be re-acquired at increased scan quality to detect lower intensity b- and y- ions to identify the peptide.

These methods introduce the concept of distinguishing linear peptide from isopeptides and the potential for chemical modifications to improve isopeptide detection. However, the power of these methods are limited to interpretation of MS2 spectra. As yet, there is no method for distinguishing isopeptides at the MS1 level.

3.1.6 Software solutions for identifying UBL isopeptides

Typical proteomic search engines are capable of dealing with simple modifications by creating an *in silico* mass shift for modified amino acids. However, they are not able to deal with composite spectra derived from fragmentation of the modifier itself. Software support for analysis of complex isopeptide fragmentation is limited. Although expected fragmentation ions can easily be calculated for predicted isopeptides, without

specialised software laborious manual interpretation of spectra is required. This limited throughput approach has been successful for sumo E3 RanBP2 autoSUMOylation sites(Cooper et al., 2005) and SUMO polymerization sites (Matic et al., 2008).

To aid manual interpretation, Matic et al. (2008) introduce the concept of a virtual modification. Because the SUMO remnant is typically longer than the substrate, isopeptides can be conceptually rotated about the isopeptide bond to make the UBL part of the main peptide (Figure 3.4). The majority of the fragment ions can therefore be accounted for by fragmenting the 'main' peptide whereas the smaller C-terminal part of the substrate peptide becomes a mass-shift without considering fragment ions.

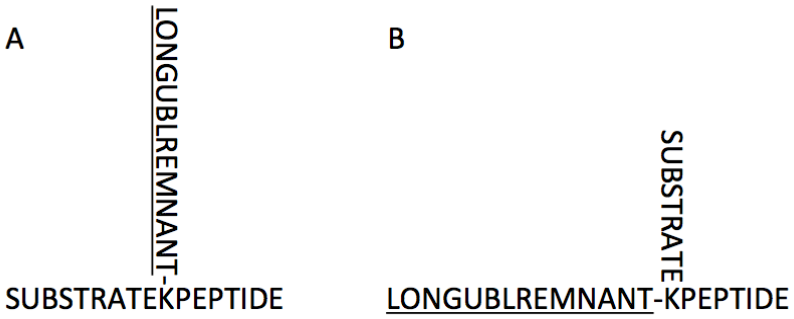


Figure 3.4 Isopeptide 'virtual' modification simplifies isopeptide spectral interpretation

An isopeptide sequence can be conceptually rotated about the isopeptide bond to make the UBL remnant part of the main peptide, and the C-terminal portion of the substrate the 'virtual' modification.

The ChopNSpice tool was developed to make composite SUMO-substrate peptide databases (Hsiao et al., 2009). This program takes the C-terminal UBL peptide remnant and all possible substrate peptides and creates linearised UBL-substrate peptide sequences. The tool was demonstrated to identify SUMO2 isopeptides *in vitro* and *in vivo* and has also been used to also identify isopeptides from FAT10 (Leng et al., 2013). In a similar strategy was used by Tammsalu et al. (2014), but using only substrates detecting using a shortened SUMO KGG mutant. Their virtual branched peptide

databases was used to identify SUMO isopeptides in existing datasets of wild-type SUMO purifications. A significant benefit of this tool is that the resulting databases can be searched by standard proteomic search engines thus making SUMO isopeptide analysis achievable without a high degree of manual interpretation. However, there can be significant loss of spectral assignment as a result of peptide linearisation. Figure 3.5 outlines how fragment ions derived from the C-terminal part of the substrate peptide can no longer be assigned. This results in diminished confirmation of the substrate peptide assignment, which is of concern when large UBL modifiers like SUMO2 already often dominate over substrate fragmentation.

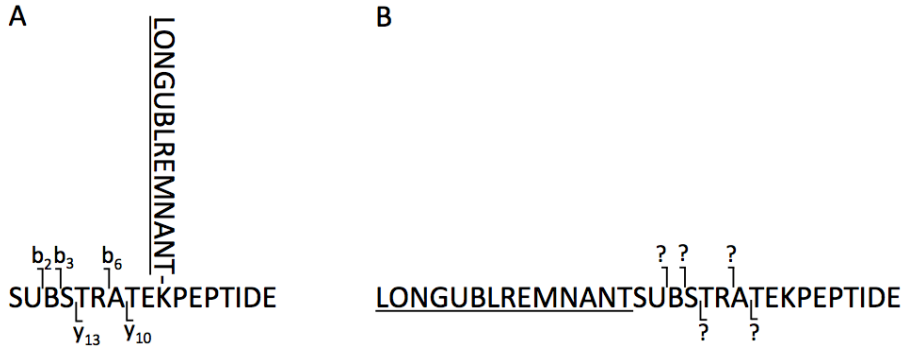


Figure 3.5 The ChopNSpice strategy results in loss of spectral assignment
True fragment ions from the C-terminal part of the substrate peptide (A) no longer correlate with predicted fragment ions after creation of the linearised peptide (B). The substrate C-terminal b- and y-ions (examples indicated) increase or decrease by the mass of the UBL, respeciviely. Ions derived from the N-terminal part of the substrate peptide and the UBL can be assigned correctly however.

The SUMmOn pattern recognition software was developed specifically to address the complex fragmentation derived from long SUMO tryptic remnants (Pedrioli et al., 2006). Because SUMO isopeptides result in fragmentation series from both the target and modifying peptides, SUMmOn performs two independent scorings for the SUMO modification and the target peptide. SUMmOn can therefore match all fragment ions from both UBL modifier and substrate peptide, unlike proteomic database searches

using chopNSpice databases. SUMmOn can also recognise and score SUMO ions in spectra for which there is no candidate substrate peptide. SUMmOn has been used to identify unexpected SUMO1 sites (Pedrioli et al., 2006), and in conjunction with a lysC digestion, NEDD8 polymerisation sites (Jeram et al., 2010).

3.1.7 Objectives of research

UBL modifications should be considered in the context of the entire proteome rather than in isolation. Global proteome analysis, and analyses of many modifications including UBLs, will ultimately become common practice when characterising biologic systems. UBL research should therefore be heading towards isopeptide analyses with integration into standard proteomic workflows. Genetic manipulations are generally not feasible requiring analysis of long remnants. A sample processing with an enzyme of high specificity is essential for complex samples, thus creating a very short remnant will not always be practical for all UBLs. Current success in the analysis of UBL isopeptides with long remnants is limited, even for *in vitro* UBL modifications. Expanding our ability to study UBLs in a broader range of biological scenarios would have significant benefits, especially for medical applications.

This research focuses on improving analytic techniques for long UBL remnants. To address the specific challenges involved, no attempts will be made to shorten isopeptide remnants. UBLs will be maintained as native sequences and trypsin will be the preferred proteolytic enzyme. Isopeptide identification will be approached from the bench, at the mass spectrometer, and bioinformatically, in an attempt to create a straightforward approach to tackling this complex domain of research. Attention is given mainly to SUMO, as the tryptic remnants from SUMO1/SUMO2 are the longest and most difficult to analyse.

3.2 Materials and Methods

3.2.1 Plasmids

Plasmids were constructed by the Dundee MRC-PPU cloning facility.

Plasmid pET156P-6HIS-SUMO1 (MRC-PPU cloning Ref: DU32080) genebank NP_003343. Protein sequence (13.6 kDa):

MGSSHHHHHHSSGLEVLFQGPGSMSDQEAKPSTEDLGDKKEGEYIKLKVIG QDSSEIHFKVKMTTHLKKLKESYCQRQGVPMNSLRFLFEGQRIADNHTPKELG MEEEDVIEVYQEQTGG

Plasmid pET156P-6HIS-SUMO2 (MRC-PPU cloning Ref: DU32081) genebank AK311837.1. Protein sequence (13.1 kDa):

MGSSHHHHHHSSGLEVLFQGPGSMADEKPKEGVKTENNDHINLKVAGQDG SVVQFKIKRHTPLSKLMKAYCERQGLSMRQIRFRFDGQPINETDTPAQLEMEDE DTIDVFQQQTGG

3.2.2 Proteins reagents

Ubiquitin and K11, K48, K63 ubiquitin dimers were produced by the Dundee MRC Protein Production unit. Recombinant RanGAP1(481-587), E1-SAE1/2 and Ubc9 proteins were kindly supplied by the R. Hay's lab, University of Dundee. Recombinant RNMT (human isoform 1) was kindly provided by V. Cowling's lab, University of Dundee.

3.2.3 Media

M9 minimal media (1L): 200 mL of M9 salts (1 L stock 64 g Na_2HPO_4-$7H_2O$, 15 g KH_2PO_4, 2.5 g NaCl, 5 g NH_4Cl), 2 mL of 1 M $MgSO_4$, 20 mL 20% glucose, 100 µL 1M $CaCl_2$

Defined auto-induction media, final concentrations: 25 mM $(NH_4)_2SO_4$, 50 mM KH_2PO_4, 50 mM Na_2HPO_4, 0.8% (v/v) glycerol, 0.015% (w/v) glucose, 0.5% (w/v) lactose, 2 mM $MgSO_4$, 20 μM $CaCl_2$, 10 μM $MnCl_2$, 10 μM $ZnSO_4$, 2 μM each of $CoCl_2$, $CuCl_2$, $NiCl_2$, Na_2MoO_4, Na_2SeO_3, and H_3BO_3, 0.2 μM each of nicotinic acid, pryridoxine, thiamine, p-aminobenzoic acid, folic acid, riboflavin, vitamin B12. 0.2 g/L of each of the 20 standard amino acids except F/W/Y/C.

3.2.4 Bacteria Transformation (E. coli DH5-alpha and BL21)

Approximately 50 ng plasmid was mixed with 100 uL of BL21 competent cells for 30 min on ice. Cells were heat shocked for 45 sec at 42 °C and then incubated on ice for 2 min. After addition of 0.9 mL of LB, cells were incubated for 1 hr at 37 °C before collecting cells by centrifugation at 3,000 x *g* for 5 min and plating on LB-amp or LB-kan plates.

3.2.5 Expression of SUMO1 and SUMO2 in M9 minimal media

BL21 cells transfected with SUMO2 plasmid were grown as precultures in M9 media with 50 μg/mL carbenicillin (Sigma)at 37 °C. Pre-cultures were diluted to an OD600 of 0.05 and grown to approximately OD600 of 0.6. Cultures were sampled (pre-induction control) then split into two cultures. For SUMO1, either tyrosine (Formedium) or $^{13}C_9{}^{15}N$-tyrosine (Cambridge Isotopes) to 0.2 mg/ml (from solid). For SUMO2, either phenylalanine (Formedium) or $^{13}C_6$-phenylalanine (Cambridge Isotopes) was added to 0.2 g/L (from a 100x stock). An additional 50 μg/mL carbenicillin was added, the temperature was dropped to 20 °C, and cells were induced with 1 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG; Sigma). Cells were grown for 19 hr (to OD600 ~2.0) before collecting cells by centrifugation at 3,000 x *g* for 5 min. Cell pellets were resuspended in cold buffer (200 mM phosphate pH 7.5, 250 mM

NaCl, 0.1% triton, 0.5 mM EGTA, 1 mM EDTA, 10 μg/mL leupeptin (Sigma), 1 mM Phenylmethylsulfonyl fluoride (PMSF, Roche), 1 mM Tris(2-carboxyethyl)phosphine (TCEP, Thermo Scientific)). Cells were ruptured using a probe sonicator (Branson) on ice for 30 sec at 30% amplitude. Sonications were repeated 3 times with 1 min rests on ice between each. Debris was pelleted by centrifugation at 3000 x *g* for 15 min. The supernatant was applied to NiNTA agarose (Quagen), detergent was removed with thorough washes in PBS and beads were changed to a clean eppendorf before elution. His-SUMO1/2 was eluted in PBS/400 mM imidazole and dialysed for 16 hr at 4 °C against PBS, 0.5 mM TCEP in a 1 kDa dialysis unit (Amersham). Note that His-tag and precission cleavage site (LEVLFQGP) remain intact on the N-terminus of proteins.

3.2.6 Expression of SUMO1 in DIAM

SUMO1 was expressed as described for M9 media expression except for the following amendments. Cells were grown in DAIM-CYWF media and was supplemented with F, W, and either light Y or heavy $^{13}C_9{}^{15}N$-Y to 0.2 mg/ml. Cells were still induced with IPTG even though the media is can auto-induce.

3.2.7 Expression of N15 SUMO2

Expression was conducted as above but with the following amendments. Cells were grown in ISOGRO ^{15}N Growth Medium (Sigma) at 98% ^{15}N isotope purity. Cells were induced with 100 μM IPTG and the temperature was dropped to 15 ºC for overnight growth. Cells were lysed in 50 mM Tris/HCl pH 7.5, 250 mM NaCl, 1% Triton, 1mM EDTA, 1mM EGTA, 1mM DTT. Cells were eluted from NiNTA in 50 mM HEPES pH

7.5, 150 mM NaCl, 0.03% Brij-35, 0.5 M imidazole. Eluted protein was buffer exchanged into waster using PD-10 Desalting Columns (GE healthcare).

3.2.8 *In vitro* SUMOylation reactions

5 µg RanGAP1 was SUMOylated with 0.5 µg E1 (SAE1/2), 1.5 µg E2 (UBC9), and 3 µg 1:1 mix SUMO2/$^{13}C_6$-phenylalanine heavy SUMO2 in 20 µL 50 mM Tris (pH 7.5), 5 mM $MgCl_2$, 2 mM ATP, and 5 mM DTT for 2 hr at 37 °C.

SUMOylated RNMT was kindly provided by Thomas Gonatopoulos Pournatzis (Cowling lab, MRC, Dundee University) but using SUMO reagents synthesised as above. In a total final volume of 10 µl, 3 µg recombinant RNMT were incubated with 0.1 µg E1-SAE enzyme, 0.75 µg Ubc9 and 3 µg 1:1 mix SUMO2/$^{13}C_6$-phenylalanine heavy SUMO2 in the presence of 50 mM Tris (pH 7.5), 5 mM MgCl2, 5 mM ATP. The samples were incubated at 37 °C for 5 hr.

SUMO reactions were terminated at 95 °C for 5 min. Protein was reduced in 0.5% RapiGest and 10 mM DTT at 37 °C for 40 min, then cysteines were alkylated with 25 mM chloroacetamide (Sigma) for 30 min. The sample was diluted 10x with 100 mM ammonium bicarbonate and digested with trypsin at 1:40 wt/wt. RapiGest was precipitated with acidification with trifluoracetic acid and centrifugation at 17,000 x *g* for 15 min, and peptides were purified on a C18 MicroSpin column (The Nest Group) and dried.

3.2.9 Isotope labelling ubiquitin isopeptides.

30 µg mixture of K11/K48/K63 ubiquitin dimers (Protein Production group, College of Life Sciences, University of Dundee) were amine blocked by four additions of either

50 mM nicotinic NHS ester (synthesised as previously described (Münchbach et al., 2000)) from a 1 M stock in dimethylformamide at 1 hr intervals at 37 °C with a final incubation time of 8 hr followed by tryptic digest and peptide clean-up on C18. Peptides were isotope labelled in 125 mM TEAB 70% (v/v) isopropanol with 1U mTRAQ Δ0 or Δ4 (Applied Biosystems). Serine, threonine, and tyrosine esters were deblocked with 5% (w/v) hydroxylamine (Sigma) for 1 hr at 37 °C followed by an additional C18 cleanup.

3.2.10 LCMS data acquisition

Peptides were injected onto a self packed 75 μm inner diameter PicoTip Emitter (New Objective), packed with Magic C18AQ 3 μm 200 Å beads (Michrom Bioresources) which was heated to 45 °C. A Proxeon EASY-nLC or a Dionex Ulitimate 3000 HPLC delivered a 300 nL/min gradient using buffers A (0.1% formic acid, 2% acetonitrile) and B (0.1% formic acid, 90% acetonitrile) for peptide analyses. Gradients were run from 1-40% B over 60 min (*in vitro* RanGAP1 SUMOylation) or 120 min (RNMT SUMOylation) followed by a high solvent wash and equilibration. Data were acquired on a Velos Orbitrap mass spectrometer (Thermo Fisher Scientific), at 100k Orbitrap resolution, and with a preview scan triggering data dependent acquisition (DDA) of the top 12 precursors above a 500 precursor intensity threshold. Peptides were isolated within a 2 m/z window for fragmentation by rapid scan ion trap CID.

3.2.11 Targeted LCMS data acquisition

Targeted LCMS analysis were conducted as for RanGAP1 analysis method but with method variations including; monoisotopic precursor selection on/off, charge state

screening on/off, dynamic exclusion on for 30 sec after 1 repeat or off, MSn trigger threshold at 500 or 5000 counts.

3.2.12 Infusion MS data acquisition of tyrosine AAs

50 μM solution of tyrosine or $^{13}C_9{}^{15}N$-tyrosine was prepared in 30% methanol, 0.1% formic acid. The solution was infused by Hamilton syringe at 4 uL/min to an electrospray source. MS acquisition was acquired at 100k resolution

3.2.13 Data analysis

LCMS raw data files were converted to mzXML format using ReAdW.exe. Assignment of SUMO isopeptide spectra was done using SUMmOn (Pedrioli et al., 2006) using precursor monoisotopic masses with 3 Da tolerance, fragment average masses at unit tolerance, charges states 3-9, and permitting oxidised methionine. Each analysis was repeated for 8 combinations of SUMO2 fully-cleaved/miss-cleaved, oxidised methionine/non-oxidised methionine, light/heavy phenylalanine.

Detection and assignment of isopeptides, and compilation of SUMmOn results, was conducted using in-house software. See Results and Discussion for details, and Appendix E for details on usage and parameters. External feature detection algorithms also used included using Progenesis (www.nonlinear.com), Maxquant (Cox and Mann, 2008), OpenMS (Kohlbacher et al., 2007), msInspect (May et al., 2009), and superHirn (Mueller et al., 2007). Viewing ICID and SUMmOn/ICID compiled data (tab delimited format) was viewed in Microsoft Excel. Manual validation of spectra was conducted in Xcalibur 2.2 Qual Browser (Thermo Scientific)

For targeted analysis in tryptic yeast, MS2 spectra were assigned to peptide identifications by searching against a *S. cerevisiae* protein database (version 2011-02-03 from SGD http://www.yeastgenome.org/) using X!!Tandem release 10-12-01-1 (MPI parallelised version of X!Tandem, http://wiki.thegpm.org/wiki/X!!Tandem). Search results were validated using the Trans-Proteomic Pipeline (TPP) v4.6 running on Linux. Peptide assignments were validated using PeptideProphet and accepted above 1% FDR.

3.2.14 Software development

All software was developled in Eclipse 4.3 (Kepler) using Java 1.6 on Mac OSX and run on both Mac OSX and CentOS linux platforms. External JAR libraries used include jopt-simple-4.3.jar (http://pholser.github.io/jopt-simple/) and jrap_StAX_v5.2.jar (http://sourceforge.net/projects/sashimi/).

3.3 Results and Discussion

3.3.1 Strategy for detecting isopeptides by isotope coding

Isotope labelling for quantitative proteomics is commonplace, such as SILAC and chemical labelling with heavy formaldehyde. For peptides that are confidently identified, isotope labelled peptide pairs can be interrogated to derive quantitative information. The strategy presented here applies the use of isotope labelling to the reverse of this concept. Instead of extracting isotopic information from identified peptides, detection of isotope labelled peptide pairs can be used as a method of identifying peptides. Specifically, peptide isotope labelling is used to uniquely distinguish UBL modified isopeptides from unmodified linear peptides.

To target UBL modified peptides, their unique features need to be exploited to distinguish them from unmodified peptides. The labelling strategy, outlined in Figure 3.6, uses an equal mixture of light and heavy stable isotope labelled UBL for *in vitro* reactions. The use of isotopes enables detection of isopeptides as distinctive isotopic pairs in high resolution MS1 scans. An important feature of this strategy is the unique ability to predetermine UBL isopeptide features within a dataset without the use of MS2 fragmentation data. This approach is therefore of particular benefit for the analysis of isopeptides with very complex fragmentation where MS2 spectral identification is difficult, most notably SUMO2 isopeptides with a 32 amino acid tryptic remnant.

The predetermination of UBL isopeptide features enables alternative strategies to be undertaken to determine the identity of the UBL conjugation site. In the simplest case, assignment of a peptide identification can be made by accurate precursor mass alone. The detection of isopeptide precursors can also complement existing isopeptide

identification strategies by adding confidence to MS2 interpretations. Having a list of candidate isopeptide features also enables a targeted MS analysis to acquire additional MSn spectra.

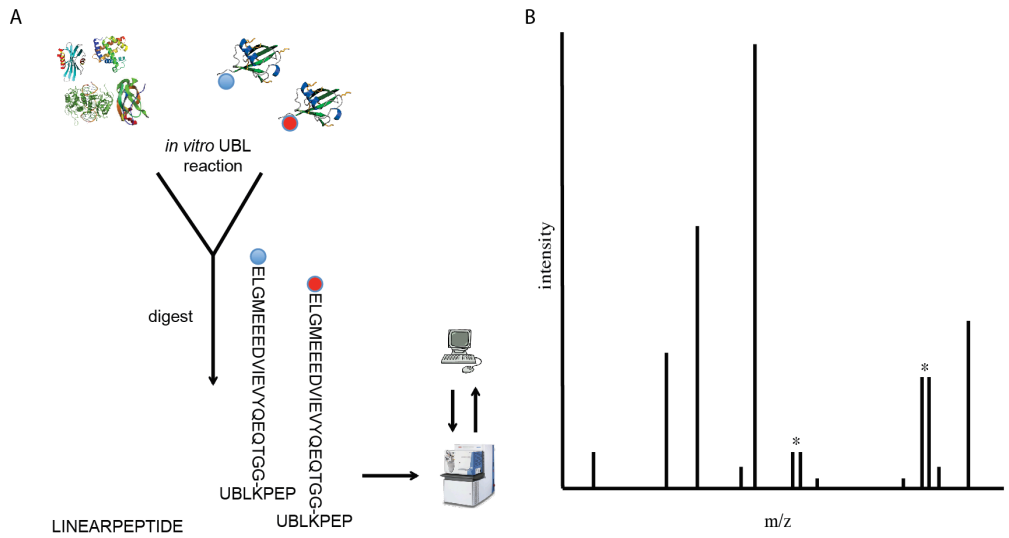


Figure 3.6 Conceptual overview of the isotope coded isopeptide detection (ICID) strategy

A) Light and heavy isotope coded UBLs are used in equal proportions for *in vitro* substrate modification. The reaction is digested to generate peptides and isopeptides prior to mass spectrometry analysis. Bioinformatic analysis is performed to detect diagnostic isotope pairs to support identification of isopeptides. B) A depiction of a spectrum containing unmodified peptides and isotope labelled isopeptides (*) which are observed as pairs of features at diagnostic mass intervals.

In this strategy, the *in vitro* reaction, sample processing and initial MS acquisition are straightforward and do not require deviation from existing experimental procedures. The novel aspects of this workflow lie in the generation of isotope labelled UBL and the bioinformatic interpretation of the data. A more detailed diagram of the workflow is presented in Figure 3.7. Both generation of isotope labelled UBL and development of bioinformatic workflows will be major focuses of this chapter.

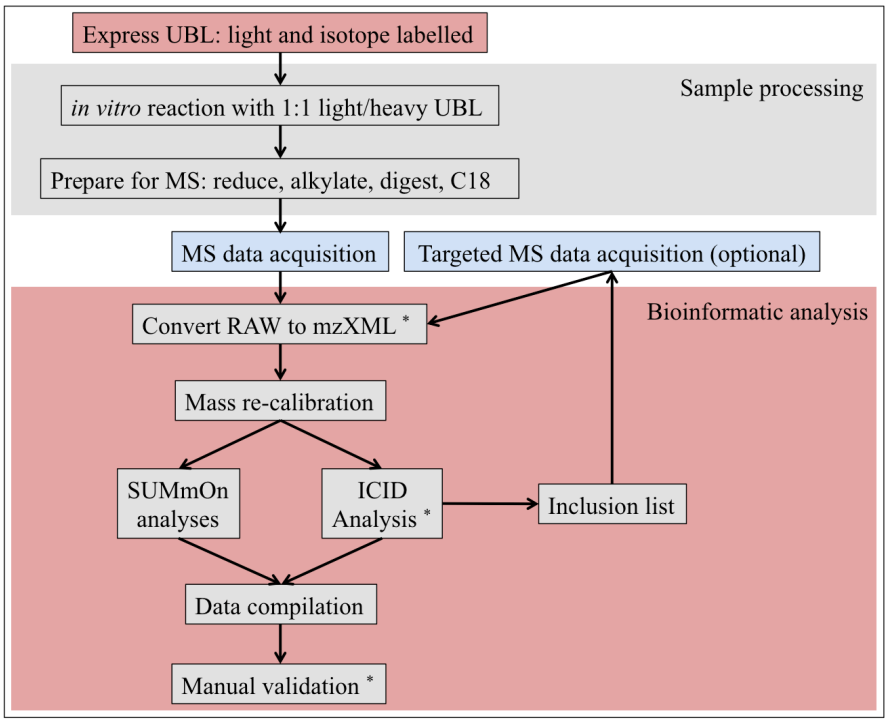


Figure 3.7 Illustration of the ICID sample processing and bioinformatic workflow

Individual steps in the ICID workflow are presented and a can be categorised into UBL expression, sample processing, MS data acquisition, and bioinformatic analysis (essential elements indicated *). Novel aspects of this work included expression of isotope coded UBL and bioinformatic interpretation of MS data.

3.3.2 Synthesis of isotope labelled UBLs

Chemical labelling of UBLs is not practical because they must remain functional for *in vitro* reactions. UBLs must therefore be isotope labelled during expression in *E. coli.* To make an effective isotope labelled UBL, expression protocols must be modified to include isotopes within the C-terminal region of the UBL. The isotopes must remain within the proteolytic remnant after digestion so that the isotopes remain associated with the substrate peptide. This could potentially be achieved by atomic-level labelling with nitrogen or carbon. Alternatively, a SILAC approach could be taken to incorporate heavy amino acids. Ideally, isotopes should not alter chromatographic elution time, so deuteriated amino acids should be avoided if possible. However, no single amino acid would necessarily be suitable because UBLs do not all share a high degree of sequence similarity. Both atomic-level and amino acid level labelling were explored.

3.3.2.1 ^{15}N labelling

As a general strategy for labelling UBLs, heavy nitrogen salts could be used to label every amino acid ensuring that the C-terminal peptide is labelled for all UBLs. SUMO2 was expressed in *E. coli* using media containing 98%-purity ^{15}N ammonium salts as a sole source of nitrogen. The SUMO2 product was digested with trypsin and the resulting digest analysed by LCMS. It was observed that using 98% ^{15}N produces heavy peptide isotopic distributions with only ~50% of the signal falling on the desired heavy monoisotope (Figure 3.8). The 32-amino acid long C-terminal tryptic peptide includes 38 independent nitrogen incorporations resulting in an accumulation of impurity. Note that a 2% ^{14}N impurity may be tolerable for SILAC where multiple isotopes are incorporated as a single amino acid. The methodology presented in this chapter will rely heavily on detection of light/heavy isotope intervals. Without a clear monoisotopic signal, detection of isotope pairs will be difficult and will result in missing isopeptide discoveries. With only ~1.5 nitrogens per amino acid, this approach might be feasible for UBLs that generate shorter tryptic remnants. However, the high isotopic purity required makes it impractical for UBLs that generate long remnants.

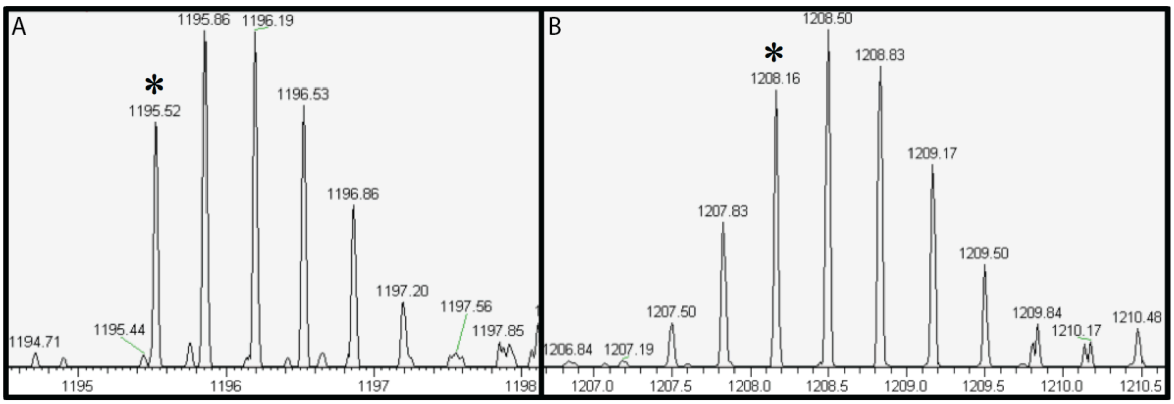


Figure 3.8 98% ^{15}N isotope labelling generates poorly defined isotope clusters.
The large SUMO2 C-terminal tryptic peptide FDGQPINETDTPAQLEMoxEDEDTIDVFQQQTGG generates a light isotope cluster with a clear monoisotopic peak (A) whereas the 98% ^{15}N heavy peptide does not have a well defined monoisotope (B). Expected monoisotopic signals are indicated (*).

3.3.2.2 SILAC labelling

Due to restrictions on isotope purity for long peptides, a SILAC approach is necessary to label UBLs. The caveat of this approach is that expression of each UBL will need to be tailored to the UBL sequence. SILAC with heavy arginine and lysine is an intuitive approach and an arginine/lysine auxotrophic *E. coli* has been used to express heavy SUMO2 (Matic et al., 2011). For the purpose of this research, an arginine and/or lysine labelled UBL is of limited use because tryptic digestion will cleave after both arginine and lysine leaving the C-terminal peptide without a heavy amino acid. It would however be feasible if using enzymes other than trypsin, such as glu-C, but only if an arginine or lysine remains on the C-terminal remnant. This would lengthen the remnant making identification even more difficult, but it would serve to generate isopeptides that would distinguish between ubiquitin, NEDD8, and ISG15. This may also be of use when substrate proteins are not suited to analysis with trypsin due to a lack of or an abundance of tryptic sites. Since isopeptides derived from ubiquitin, NEDD8, and ISG15 are not difficult to identify under most circumstance, attention was given to more challenging UBLs such as SUMO.

Labelling with heavy amino acids without generating additional auxotrophs was preferred so that the methods might be extrapolated to a general method for expressing heavy UBLs. The aromatic amino acids were selected for the option to inhibit their synthesis by blocking the shikimate pathway with glyphosate (N-(phosphonomethyl)glycine, a herbicide), which has been applied to incorporation of unnatural aromatic amino acids (Neerathilingam and Markley, 2010). Isotope coding was implemented using $^{13}C_9{}^{15}N$-tyrosine and $^{13}C_6$-phenylalanine for SUMO1 and SUMO2, respectively. This would result in a single tyrosine incorporation into the C-

terminal SUMO1 peptide and 2 phenylalanines in the C-terminal SUMO2 peptide. In the case of a missed-cleavage, the SUMO2 peptide will include 3 phenylalanine residues (refer to Figure 3.1 for sequences). Note that light SUMO1 and SUMO2 also need to be mixed at 1:1 after expression. Mixing the isotopes at 1:1 during expression can only be done if there is exactly one amino acid being incorporated, otherwise a distribution of isotopes will be incorporated.

3.3.2.3 SILAC labelling SUMO1 and SUMO2 in minimal media

Using the prototrophic BL21 *E. coli* strain, SUMO proteins were expressed in minimal media with tyrosine or phenylalanine supplementation, but without glyphosate treatment (Figure 3.9). In-gel tryptic digestion and LCMS analysis of $^{13}C_6$-phenylalanine SUMO2 identified the expected heavy peptide and no light peptide was observed (Figure 3.10). The phenylalanine in the media was therefore the sole source of phenylalanine indicating that feedback inhibition of phenylalanine synthesis must have occurred. This observation is consistent with expected inhibition of pheA (Nelms et al., 1992). No heavy isotopes were detected above natural isotope distributions in any other amino acids indicating that phenylalanine isotopes were not being recycled into other amino acids.

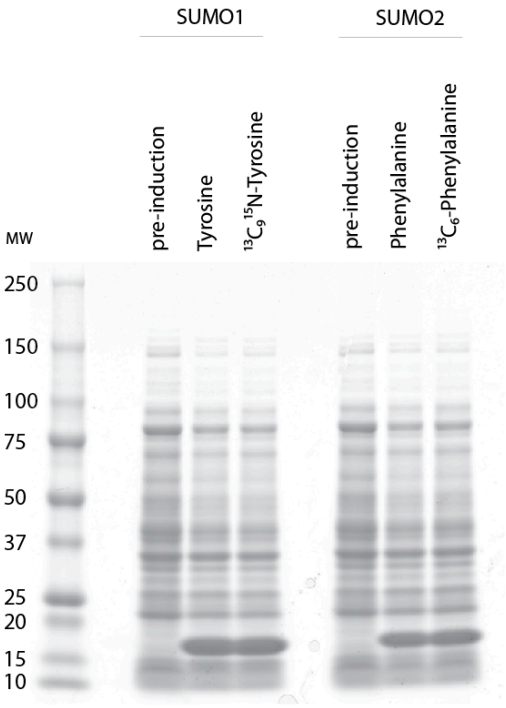


Figure 3.9 Expression of heavy SUMO1 and SUMO2 in minimal media

SUMO proteins were expressed using in minimal media (without amino acids). Expression cultures were supplemented with either light or heavy tyrosine for SUMO1 or light or heavy phenylalanine for SUMO2. A comparison between non-induced and induced cultures indicate a high level of expression of both SUMO1 and SUMO2 (~17kDa) under these growth conditions.

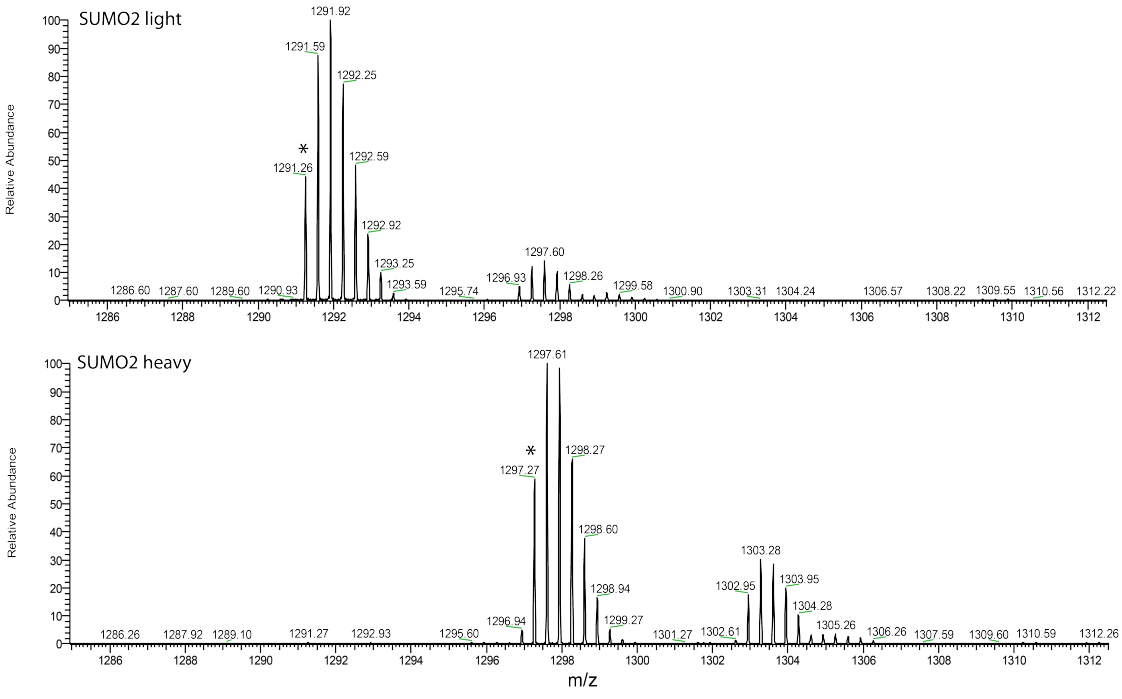


Figure 3.10 Heavy phenylalanine incorporation into SUMO2 is complete in phenylalanine supplemented minimal media

MS spectra of the C-terminal peptide FRFDGQPINETDTPAQLEMEDEDTIDVFQQQTGG from the light (A) and $^{13}C_6$-phenylalanine heavy (B) 6HIS-SUMO2 gel bands from Figure 3.9. The expected monoisotopic m/z from light and heavy SUMO2 are indicated (*) and correspond to an interval of 18.0604 Da (three $^{13}C_6$-phenylalanine amino acids are expected in this peptide). No light isotope is observed in the heavy expressed SUMO2 confirming isotope labelling. The isotopic clusters to the right belong to an unrelated peptide.

In contrast to the successful results observed from SUMO2 expression, the SUMO1 C-terminal peptide did not result in a clean incorporation of tyrosine (Figure 3.11). The isotopes were shifted to a wider mass distribution than the light peptide, which would not be expected from just a single heavy tyrosine. This indicates that not only are cells degrading tyrosine, but the tyrosine breakdown products are also being incorporated into other amino acids. It is noteworthy that no light tyrosine peptide was observed indicating that feedback inhibition of tyrosine synthesis was still effective.

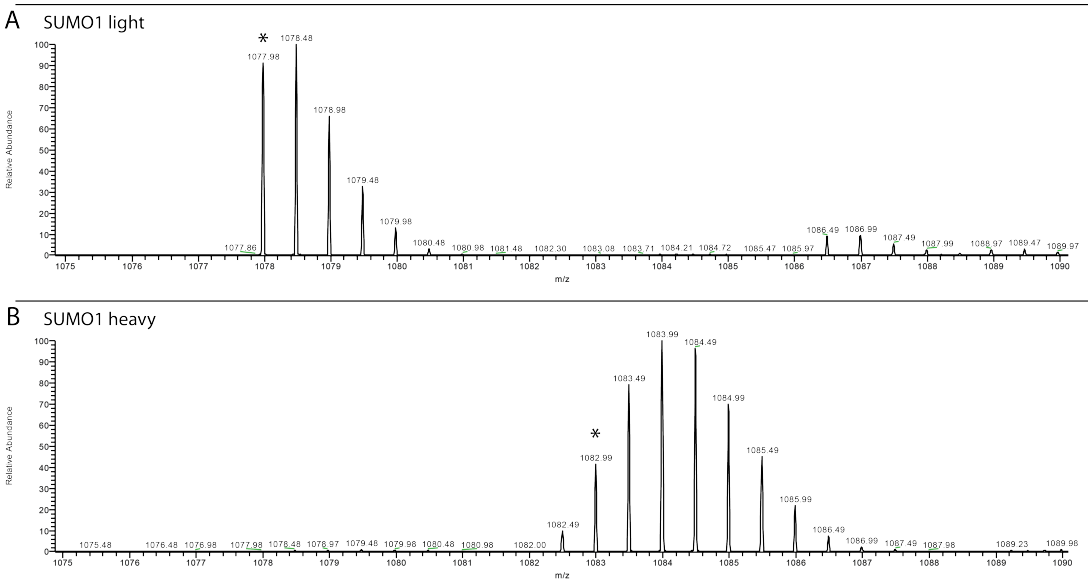


Figure 3.11 Tyrosine isotopes are recycled in minimal media

SUMO1 labelling is complete in tyrosine supplemented minimal media as indicated by the absence of light peptide. The isotope distribution of the labelled peptide is shifted to a higher mass distribution than the light peptide indicating that $^{13}C_9{}^{15}N$-Tyrosine isotopes are recycled into other amino acids. Asterisks indicate expected monoisotopic peaks from the light (upper panel) and heavy (lower panel) C-terminal SUMO1 peptide, ELGMEEEDVIEVYQEQTGG.

3.3.2.4 SILAC labelling SUMO1 in defined media

SUMO1 was expressed in defined media including amino acids in an attempt to inhibit the synthesis of all amino acids that might incorporate recycled tyrosine isotopes. In this media, heavy tyrosine containing peptides did not display an isotope distribution shifted to a higher mass range (Figure 3.12). Unexpectedly, the apparent mono-isotope was one neutron lighter than the calculated mono-isotope. MS analysis of the tyrosine used in the media confirms that the amino acid product is correctly isotope coded (Figure 3.13). The specific recycling of ^{15}N and/or 13C can be explained by metabolism occurring in the pathway after the point of feedback inhibition occurring on tyrA (Bongaerts et al., 2001). The inhibition occurs prior to the synthesis of the tyrosine precursor, hydroxyphenylpyruvate, which can be made from prephenate or by deamination of tyrosine (Figure 3.14). Re-amination of heavy hydroxyphenylpyruvate with light nitrogen can produce the lighter isotope of tyrosine that is observed. The degradation products can obviously be utilised in general amino acid synthesis, as was observed when attempting to express SUMO1 in minimal media. Phenylalanine is regulated in an analogous fashion to tyrosine but isotopes were not incorporated elsewhere. This suggests that it is the ^{15}N that is being recycled rather than the hydroxyphenylpyruvate because the phenylalanine used to label SUMO2 did not include and ^{15}N. The observed monoisotope in the current SUMO1 expression product is approximately equivalent to $^{13}C_9$-tyrosine incorporation. SUMO1 should in future be expressed with $^{13}C_9$-tyrosine to prevent isotope recycling through deamination/amination of tyrosine and to maintain a normal isotopic cluster.

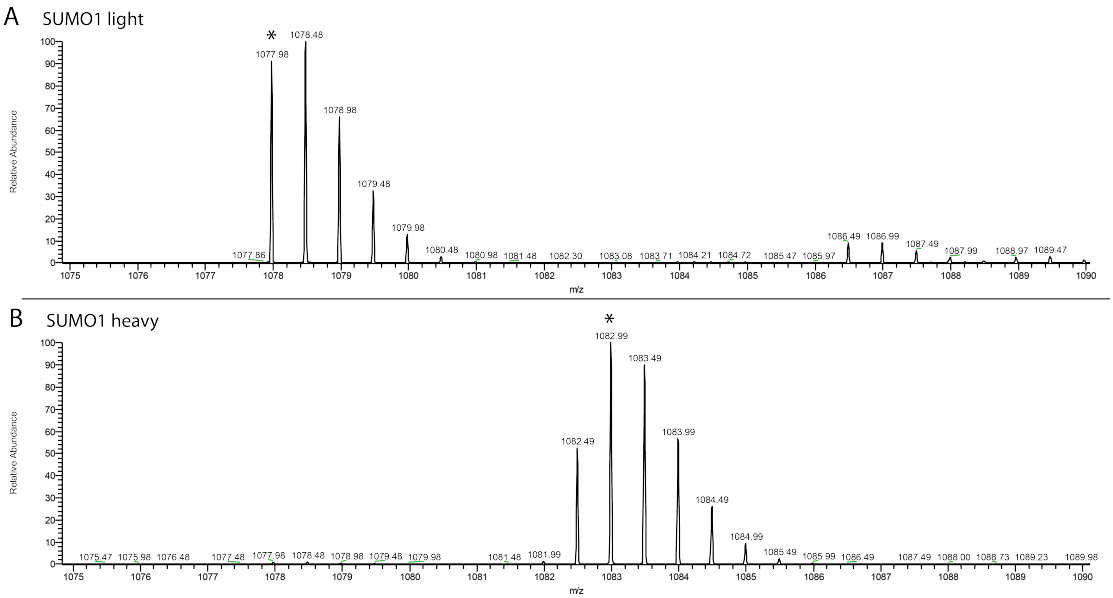


Figure 3.12 Tyrosine deamination causes loss of ^{15}N in defined media

MS spectra of the C-terminal SUMO1 peptide expressed in defined media containing light tyrosine (A) or heavy $^{13}C_9{}^{15}N$-Tyrosine (B) shows that there is a loss of an isotope from the heavy amino acid. Additional isotope incorporation appears to be inhibited by including a full supplement of amino acids (comparing to Figure 3.10B). Asterisks indicate expected monoisotopic peaks from light (upper panel) and heavy (lower panel) SUMO1 C-terminal peptide, ELGMEEEDVIEVYQEQTGG.

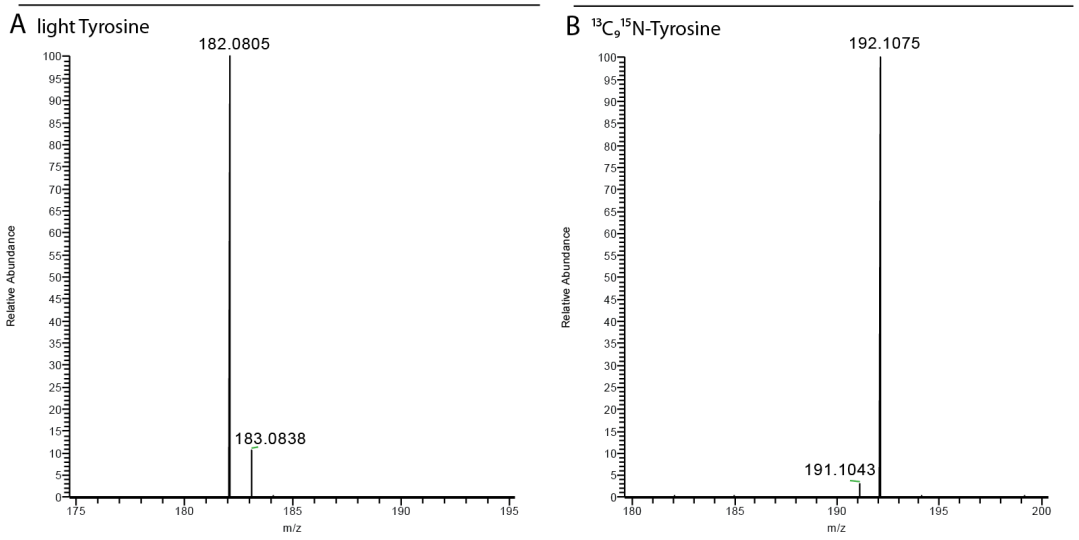


Figure 3.13 Analysis of light and heavy $^{13}C_9{}^{15}N$-Tyrosine amino acid confirms expected isotope composition and purity

Infusion MS analysis of (A) light tyrosine (Formedium) and (B) $^{13}C_9{}^{15}N$-Tyrosine (Cambridge Isotopes) used in media preparation verifies a mass difference of 10.0270 which is very close to the expected 10.0272 (~1.1 ppm error). This confirms that the unexpected light peptides observed in *E. coli* expressed proteins are a consequence of tyrosine metabolism.

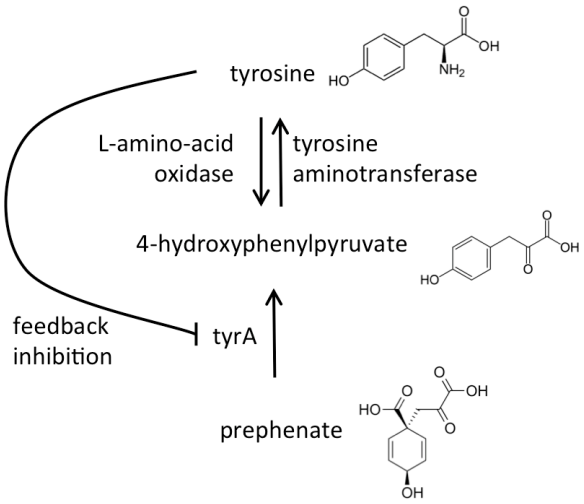


Figure 3.14 Key pathway events in tyrosine metabolism permitting isotope recycling

Recycling of the tyrosine amine group in *E. coli* can occur despite feedback inhibition of tyrosine synthesis. The amination/deamination cycle results in removal of heavy nitrogen for potential use in other amino acids and also modifies heavy tyrosine with a light nitrogen.

3.3.2.5 General labelling strategies for UBLs

Inhibition of aromatic amino acid synthesis with glyphosate turned out to be unnecessary as isotopes of natural amino acids, unlike unnatural analogues, can cause the feedback inhibition required for labelling proteins in prototrophic *E. coli*. Labelling in media with a full amino acid complement and selective replacement with heavy amino acids without α-amino isototopes appears to be an ideal general strategy for the aromatic amino acids phenylalaninine and tyrosine. Alternative strategies are apparent allowing for most UBLs to be labelled in a similar manner. In addition to the SUMO proteins, labelling with phenylalanine can also isotopically label the FAT10 tryptic remnant. Atomic labelling would work for short remnants only, such as ^{15}N labelling for Smt3 and nitrogen plus carbon for very the very short remnants from ubiquitin, ISG15, NEDD8, and Ufm1. Additional strategies are available if using proteolytic enzymes other than trypsin. Other amino acids have been reported to effectively

incorporate without needing auxotrophic cells, including histidine, lysine, methionine and alanine (O'Grady et al., 2012). A wide range of BL21 auxotrophs have also been published offering an alternative approach where a specific UBL and proteolytic enzyme required them (Lin et al., 2011; Matic et al., 2011; Mehlhorn et al., 2013).

3.3.3 ICID bioinformatic analysis

Bioinformatic analysis includes multiple steps for data processing and identification of isopeptides (refer to Figure 3.7 for a diagrammatic overview of the process). The essential steps required are only data format conversion to mzXML and execution of the ICID software, although manual validation of the results is also recommended. The ICID software is the primary focus of the bioinformatic interpretation as this is the core detection of isotope coded UBLs. Optional elements are also available, including incorporation of the SUMmOn search engine to complement this workflow.

3.3.3.1 MS acquisition and data pre-processing

A typical data dependent MS acquisition is a suitable dataset for ICID analysis. The ICID software relies on high mass accuracy thus limiting the supported MS platforms to high resolution instruments, such as Orbitraps. MS2 data are not utilised for detection of isotope coded peptide pairs therefore only high resolution MS1 spectra are required. MS method design will however impact the use of SUMmOn or any other software that interprets MS2 spectra. Alternative MS2 data acquisition methods such as targeted MS analyses are equally suitable for ICID analysis as long as high resolution MS1 scans are included in the analysis.

All data is first converted to mzXML format, which is a generic format for representing MS data (Pedrioli et al., 2004), and enables the workflow to be MS platform independent. Both the ICID software and the SUMmOn search engine

(Pedrioli et al., 2006) utilise mzXML format. Given the high dependency on MS1 data for feature detection, a well calibrated analysis with accurate precursor masses will benefit the analysis. Checking the mass accuracy of the analysis and, if necessary, performing post acquisition mass calibration is recommended. Post acquisition LCMS calibration software has also been developed (see HyperProphet chapter).

3.3.3.2 ICID software overview

The ICID software was developed in the Java language and has a simple usage requiring a single call on the terminal. A diagrammatic overview of the user interaction is presented in Figure 3.15. User defined parameters influence isotope coded isopeptide detection, assignment of peptide identifications, and inclusion list generation. Details on code usage and help on input parameters can be found in Appendix B. The ICID process includes the detection of isotope coded feature pairs, assignment of peptide identifications to the detected features, and output of results for further analysis.

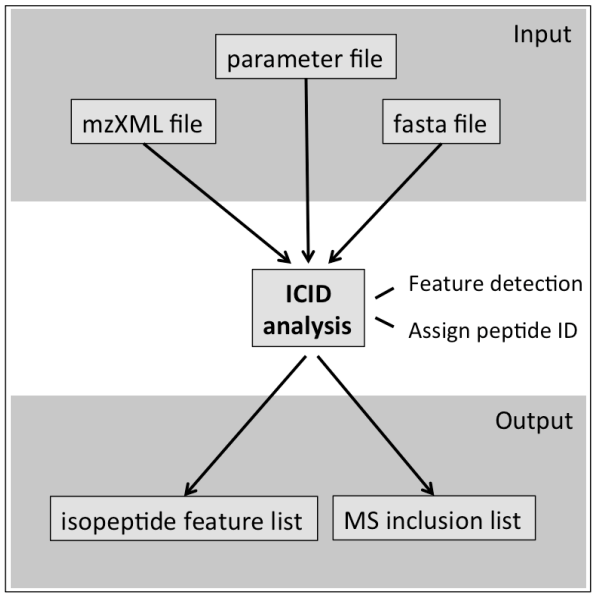


Figure 3.15 User interaction overview for the ICID software

An LCMS analysis (in mzXML format) and a protein database (in fasta format) are used as input to detect isotope coded isopeptides and assign peptide identifications under the direction of user defined parameters. The detected candidate isopeptides are output to an inclusion list (for optional targeted MS analysis) and with full supporting information for further analysis.

3.3.3.3 ICID Isotope Peak Picking algorithm

The most critical aspect of this workflow is to be able to detect isotope pairs within MS data. Initially, the feature detection algorithm from Progenesis (www.nonlinear.com) was used to detect all features in a run. Feature detection deconvoluted each peptide feature to a discrete monoisotopic m/z, with a chromatographic elution time and peak intensity. This feature set was then imported as input to determine co-eluting feature pairs at the diagnostic isotope mass interval. While this successfully detected many isotope coded peptide pairs, a limitation in isopeptide detection was observed due to errors in the initial detection of features, which were frequently missing or incorrectly assigned to the wrong isotope (Figure 3.16).

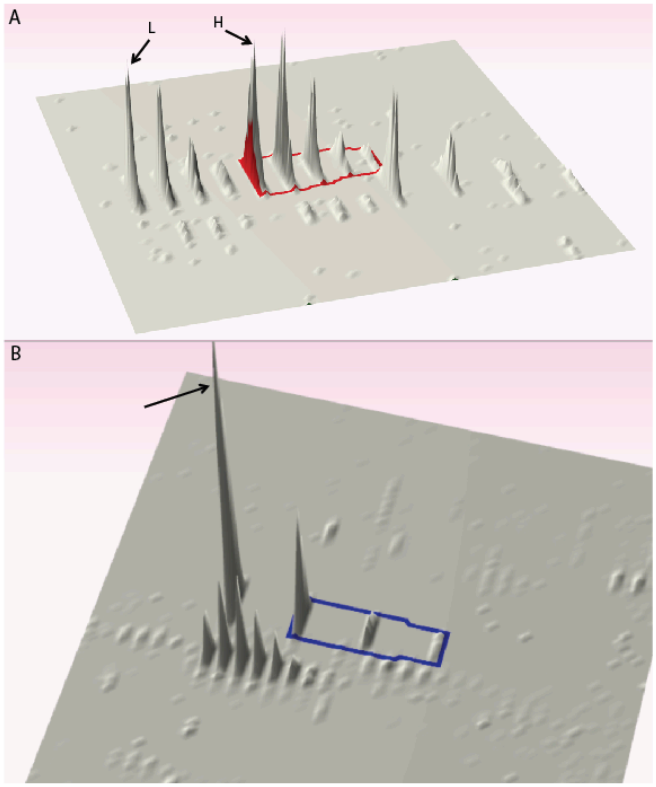


Figure 3.16 Example of erroneous feature detection in third party feature detection software

Image captures from Progenesis graphical interface displays an example of a correctly detected peptide feature of a light (L) and heavy (H) Δ4Da isotope labelled pair (A) and a feature detection error where the second isotope has been incorrectly identified as the monoisotope (the correct monoisotope is indicated with an arrow).

In an attempt to find an improved algorithm, five popular commercial and free feature detection software packages were tested, including Progenesis, Maxquant (Cox and Mann, 2008), OpenMS (Kohlbacher et al., 2007), msInspect (May et al., 2009), and superHirn (Mueller et al., 2007). A test sample was generated by mTRAQ labelling tryptic ubiquitin dimer and spiking into tryptic whole yeast lysate. Ten manually validated low abundance features, belonging to 5 pairs of isotopically labelled peptides, were used as positive controls. Of the software tested, none could successfully detect all manually validated features. As the bioinformatic process is heavily reliant on correct detection of isotope coded features pairs, code was developed to gain more control over feature detection sensitivity, improve detection of isotope pairs, and circumvent the need for third party feature detection.

The new algorithm for detecting isotope labelled peptide pairs is not a re-implementation of traditional feature detection. The approach was to over-predict all isotope pairs in raw convoluted data and subsequently validate the quality of the detected features. This relatively naive approach is practical only because the isotope feature pairs have a strict relationship that are not inherent to typical features. Isotope feature pairs are required to be co-eluting during C18 chromatography, of similar intensity, and a predetermined mass interval as defined by the isotopes used during UBL labelling. The in-spectrum mass accuracy, the mass error between peaks in the same spectra, is generally smaller than the deviation observed between spectra making in-spectrum isotope detection a discerning parameter. These strict criteria permitted the omission of traditional feature detection and isotope pairs could be extracted directly from raw data.

The main steps in the algorithm are as follows:

- Import mzXML file (in centroid format)
- Find local maxima over m/z-time domains (over-prediction of chromatographic peaks).
- Find all mass pairs at expected isotope interval(s), restricting isotope pairs to co-eluting features and exploiting high in-spectrum mass accuracy.
- Confirm each match feature is the monoisotope of an isotopic cluster.
- Confirm each feature pair is of the expected charge state.
- Confirm each feature pair is of approximately equal intensity.
- Merge duplicate/redundant features.

The new algorithm proved to be more sensitive and successfully detected all the manually validated feature pairs (Table 3.1). The validation of feature quality is more relaxed than traditional feature detection and retains features unless they clearly fail to meet the minimal criteria. As a result of increased sensitivity, the new algorithm resulted in a higher number of predicted isotope features detected than most other algorithms. These predicted features will also include many false positives as the test sample is predominantly non-isotope labelled yeast peptides. However, the total number of features is still within a tolerable range for the ICID application as this is an achievable number for targeted MS analyses. Fewer false positive assignments would be expected for *in vitro* samples which will be significantly less complex than whole cell proteome. The algorithm was also significantly faster than the third party feature detection and also streamlined the process by removing requirements for external software.

	msInspect	openMS	superHirn	Progenesis	maxQuant	**ICID software**
total input features	320290	44017	26360	19743	79909	**553729**
manually validated features found /10	5	5	5	6	8	**10**
total isotope coded feature pairs found	636	59	9	51	164	**328**

Table 3.1 A comparison between feature detection algorithms

Five commercial and free feature detection algorithms were compared to the isotope feature pair detection algorithm developed for the ICID software. A benchmark against 10 manually validated isotope coded features indicates that the ICID software is more sensitive at performing this task than all other algorithms tested. Note that not all detected feature pairs are necessarily correct.

3.3.3.4 Assignment of peptide identifications

Detection of UBL isopeptide candidates via isotope coding allows classification of chromatographic features by their UBL modifying group. If a detected peptide is in fact a UBL modified isopeptide then the mass of the substrate peptide can be inferred by subtraction of the UBL tryptic remnant. The substrate peptide identification can, at least in low complexity *in vitro* protein mixtures, be assigned solely based on high precursor mass accuracy. Mass accuracy is therefore critical during peptide assignment as it is based solely on MS1 precursor mass without consideration of MS2 spectral identification. In this respect, identification is not unlike peptide mass fingerprinting (PMF) which has been a very common identification technique for MS instruments without MS2 capability (Pappin et al., 1993). This approach should be contrasted to precursor mass identification assignment without the use of isotope coding. If seeking a match in an all-by-all search space between all predicted peptides and all chromatographic features, then there is a high probability of finding a false positive match. By reducing the search space to a smaller set of high confidence features, assignments can be applied with higher certainty.

The ICID software accepts a fasta protein database with which to assign isopeptide identifications. The fasta database should contain only proteins know to be in the *in vitro* reaction to minimise false positives. Detection of any unexpected contaminants can be achieved using a conventional proteomics search engine. Protein sequences should ideally be manually curated to include any sequence processing and sequence modifications, such as N-terminal affinity tags. The fasta database is digested *in silico* with a custom enzyme model. Default parameters enforce a missed cleavage around a lysine to allow for the UBL attachment. The N-terminal peptide is also retained without

and internal lysine to allow for N-terminal UBL modification. Additional missed cleavages are permitted within two residues from the N- or C- terminus without penalty to account for trypsin being an endo-protease.

3.3.3.5 ICID output and manual validation

ICID results are presented as a simple tabular format. Key information exported includes;

- Calculated substrate peptide MW, and the assigned peptide/protein identification
- in-spectrum mass accuracy for the detected isotope feature pair
- Intensity of each feature and deviation from expected 1:1 ratio
- Feature parameters: m/z values, MW, charge, elution time, scan number

Considering the inherent difficulty in isopeptide identification, the data presentation is relatively simple and similar to what might be exported from a conventional proteomic search engine. Results are pre-filtered by the software based on user parameters, but data can be further viewed, sorted, and filtered using any preferred spreadsheet software or scripting environment such as R. The most confident set of identifications can be selected by considering basic criteria, most of which are common to validation of standard proteomics search results. Confidently identified peptide assignments should be supported by small ppm errors on both the isotope feature pair interval and peptide assignment. Multiple entries covering a peptide sequence also gives support for identifications due to observing multiple charge states and missed cleavages. Missed cleavage of the SUMO2 remnant is also frequently observed due to the close proximity of two arginines. There are no MS2 spectra to manually validate, however, the MS1 spectra should be visually inspected to confirm the correct detection of light and heavy isotopic clusters. This can be done in MS vendor data viewer (e.g. Xcalibur for Thermo Orbitrap instruments) and a scan number is provided with which to

quickly find the chromatographic apices of the detected peptide features. In the case of a peptide containing multiple lysine, the actual conjugation site remains ambiguous. Interpretation of MS2 spectra is required to clarify such cases.

Many features detected by ICID remain unidentified. Non-peptide substrates will also be expected, such as a substrate mass of 18.015 Da (water), corresponding to the unconjugated C-terminal peptide of the UBL, which serves as a useful positive control. A modified mass of 121.074 Da is also often observed as a result of modifying Tris buffer. It should be stressed that detected isotope coded features without a peptide assignment may occur for genuine isopeptides due to incorrect prior assumptions of the proteins present or their sequences. Multiple instances of a calculated substrate mass for various charge states and UBL missed cleavages is indicative of a potentially genuine isopeptide but on an unanticipated peptide.

3.3.3.6 Generation of inclusion list

Due to the unique application of isotope coding, isopeptide detection is not dependent on protein sequence assumptions. Assignment of peptide identifications is optional and can be omitted in cases where only an inclusion list is required. A list of isopeptide features can be generated quickly for immediate LCMS re-analysis. This is achieved without user supervision and functions based solely only on the predetermined isotope specifications in the parameter file. An inclusion list is generated with target mass and a time range, with overlapping entries being merged, and table output conforming to Orbitrap software requirements. Masses for the light, heavy, or both features of an isotope pair can be selected for retargeting.

While many spectra can be acquired during an initial DDA acquisition, a targeted analysis allows for additional spectral acquisition where target spectra were previously missing or of poor quality. More importantly, the reduced workload of a small

inclusion list allows for alternative MS methods to be utilised. With a reduced duty cycle, the more time consuming MS2 techniques become more feasible, such as increased fill volumes, increased averages, and MS2 in the Orbitrap rather than the ion trap. Orbitrap spectra provide superior resolution but are much slower and less sensitive than ion trap CID in hybrid instruments like Orbitraps. MS3 fragmentation is also an appealing option as large SUMO remnants often dominate over target peptide fragmentation. An MS3 spectra may provide target peptide ions to validate the substrate peptide, however, this currently requires manual validation as there is no software support for this type of analysis.

3.3.4 SUMmOn analyses and data compilation

SUMmOn offers an alternative approach to isopeptide identification. Although SUMmOn analyses are not strictly required, it is a valuable resource and a complementary MS2-based strategy can add further confidence to isopeptide identifications. In contrast to the ICID approach, which relies on high resolution MS1 data, SUMmOn has a strong focus on MS2 spectral analysis to identify SUMO isopeptides. However, SUMmOn was developed for low resolution ion trap data and does not take advantage of high resolution precursor mass accuracy. MS2 spectra for large high charge state isopeptides are often acquired on dominant isotopes rather than the monoisotopes during DDA MS acquisition. Searches must therefore be done with a wide precursor error tolerance which increases false positive assignments. Careful manual validation of low resolution assignments are required, particularly for the frequent occurrences of ambiguous target peptide assignment. This is further complicated by the fact that SUMO fragmentation ions often dominate over the target peptide ions making manual validation difficult. SUMmOn also presents a logistical limitation when analysing isotope coded SUMO2. The SUMO2 tryptic remnant is often

observed with a missed-cleavage and oxidised methionine. With additional isotope coding there are eight observed SUMO2 remnant variants, each requiring independent SUMmOn analyses which must be manually validated independently.

To further extend the capabilities for SUMO isopeptide identification, a data compilation tool was developed to consolidate SUMmOn analyses and ICID results into a single analysis (Figure 3.17). Data compilation aims to achieve multiple tasks during consolidation of SUMmOn and ICID data. High resolution precursor mass filters and score thresholds are applied on SUMmOn results to reduce false positive assignment, multiple SUMmOn analyses on each UBL variant are combined into a single analysis, and results are cross correlated to ICID results to ascertain whether the SUMmOn identification is supported by isotope coded feature pair.

SUMmOn accepts the same mzXML file and fasta database as ICID as well as an additional conditions.xml file defining search parameters. A sumo.xml file is output containing search results, which is used by the compilation tool, along with additional files enabling results to be viewed via a web browser. Details on running SUMmOn can be obtained from the original publication (Pedrioli et al., 2006) and from the website (http://summon.sourceforge.net). Usage details and parameters for the compilation tool can be found at the end of Appendix B.

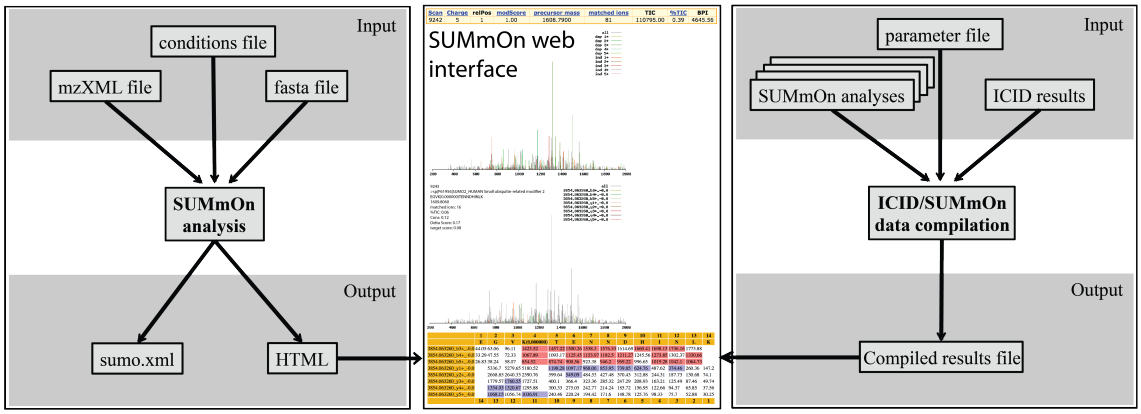


Figure 3.17 User interaction with SUMmOn and the SUMmOn/ICID data compilation tool

The mzXML and fasta input files (same as those used by the ICID analysis) are independently used by SUMmOn and ICID analyses. A conditions.xml parameter file directs the SUMmOn search engine to assign modifier and target peptide scores which are written to a sumo.xml file. Peptide assignments can be viewed through a web browser.

Results are presented in a simple tabular format for inspection in any spreadsheet software. Key information exported for each entry includes:

- SUMmOn search ID (to identify which UBL variant was searched)
- Matching ICID spectrum number for a light or heavy isotope feature (if any)
- Embedded hyperlinks to modifier and candidate peptide spectra (SUMmOn web interface)
- Precursor metrics (scan number, m/z, charge)
- Modifier and candidate peptide scores
- Candidate match peptide/protein
- Candidate match metrics (MW, ppm error, which isotope matched)

Compiled results are pre-filtered by accurate mass around the monoisotope or a larger isotope. Observed precursors around the fourth isotope are not uncommon. By further restricting results to those with both high SUMmOn scores and support from ICID, a small high quality dataset can be easily generated thus significantly reducing manual validation requirements. While the ICID analysis alone can determine isopeptide identities, the incorporation of SUMmOn now links in MS2 spectral validation. The compilation tool is SUMmOn centric and can also be run without ICID result making it a useful extension for SUMmOn analyses alone.

3.3.5 Validation on known SUMOylation substrate: RanGAP1

Ran GTPase-activating protein 1 (RanGAP1) was the first SUMO substrate identified and is known to be modified by SUMO at K524 *in vivo* (K526 for the mouse

orthologue) (Matunis et al., 1998). RanGAP1 has a preference for SUMO1 *in vivo* but is equally modified by SUMO1 and SUMO2 *in vitro* (Zhu et al., 2009). The K524 modification occurs within the ΨKXE/D consensus motif (LKSE) and modification occurs through direct interaction with the E2, UBC9. An E3 ligase is not required for efficient modification. The RanGAP1 SUMOylation reaction is well characterised and is a common positive control for SUMOylation reactions (Flotho et al., 2012) and is an ideal case study to validate the use of isotope coded SUMO for identifying conjugation sites.

RanGAP1 (C-terminal fragment 418-587) was SUMOylated with either SUMO2 or a mixture of light and isotopically heavy SUMO2. SDS-PAGE of reaction products indicate that RanGAP1 was completely modified (Figure 3.18). Both the light RanGAP1 and light/heavy SUMOylation reactions were subjected to in-solution digestion and analysed by LCMS. The resulting LCMS files were processed through the ICID software using the parameter file presented in Appendix B and a fasta file containing proteins used in the *in vitro* reaction (SAE1, SAE2, UBC9, SUMO2, and RanGAP1). The process was automated other than to filter the final result to only peptide sequences observed as multiple variants (charge state, UBL miss-cleavage, UBL oxidation). The resulting analysis (Table 3.2) successfully identified the expected RanGAP1 SUMOylated peptide encompassing K524 (LLVHMGLLKSEDK). An additional missed cleaved peptide was also identified covering the same RanGAP1 modification site (LLVHMGLLKSEDKVK). There are two lysines in the miss-cleaved peptide so the K524 modification in this peptide is presumed but not supported by the data. An additional unreported RanGAP1 SUMOylation site was discovered at K452 (LGPKSSVLIAQQTDTSDPEK). The K11 SUMO2-SUMO2 isopeptide was also detected, which is a know site for SUMO2 polymerisation and also conforms to the consensus motif (VKTE) recognised by UBC9 (Tatham et al., 2001).

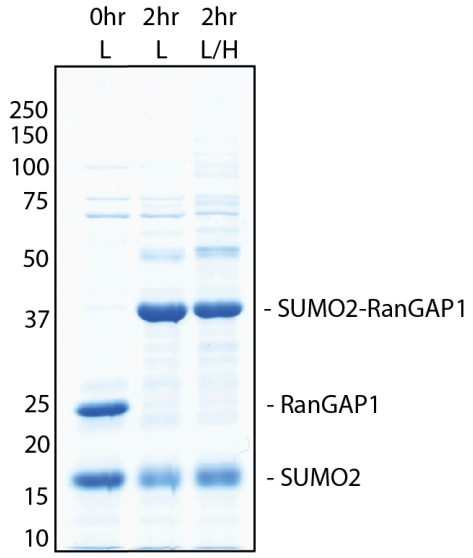


Figure 3.18 RanGAP1 is efficiently SUMOylated *in vitro* using isotopically coded light/heavy SUMO2

UBC9 efficiently SUMOylates the RanGAP1 with both light (unlabelled) SUMO2 and a 1:1 mixture of light and heavy phenylalanine labelled SUMO2. RanGAP1 is completely modified, as indicated by the loss of the unmodified RanGAP1 observed at 0 hr, and is predominantly mono-SUMOylated.

substrate MW	m/z light	m/z heavy	dMass	dMass ppm	z	intensity (Light)	dIntensity	peptide	ppm Error	protein
1609.815	1093.5115	1097.123	18.0579	-0.46	5	1757353	-1.36	EGVKTENNDHINLK	1.64	SUMO2
1609.8156	1366.6377	1371.1501	18.0498	0.21	4	483387	-1.47	EGVKTENNDHINLK	1.77	SUMO2
1609.8207	911.4284	914.4338	18.0326	-0.33	6	69314	-1.68	EGVKTENNDHINLK	2.7	SUMO2
1609.821	1032.8788	1035.2864	12.038	-0.44	5	35542	-1.39	EGVKTENNDHINLK	2.92	SUMO2
1609.8333	1290.8497	1293.8573	12.0303	-1.94	4	816219	-1.41	EGVKTENNDHINLK	5.3	SUMO2
1481.8416	1007.2829	1009.6898	12.0343	1.18	5	62935	1.3	LLVHMGLLKSEDK	2.79	RanGap1
1481.8449	1334.645	1339.1606	18.0625	0.4	4	324125	1.41	LLVHMGLLKSEDK	3.26	RanGap1
1708.9987	1391.4335	1395.9506	18.0684	0.82	4	215703	1.43	LLVHMGLLKSEDKVK	1.41	RanGap1
1709.0058	795.5375	798.1172	18.0576	-0.51	7	69323	1.8	LLVHMGLLKSEDKVK	2.67	RanGap1
1709.0099	1113.3505	1116.9613	18.0542	-1.11	5	1045959	1.27	LLVHMGLLKSEDKVK	3.41	RanGap1
1709.0184	1052.7183	1055.1246	12.0319	-1.6	5	737401	1.28	LLVHMGLLKSEDKVK	5.23	RanGap1
2113.0847	1492.455	1496.9712	18.0649	0.76	4	35495	-1.12	LGPKSSVLIAQQTDTSDPEK	-0.93	RanGap1
2113.1043	1416.6675	1419.6771	12.0386	-0.3	4	26264	-1.14	LGPKSSVLIAQQTDTSDPEK	2.48	RanGap1
18.0265	1190.1949	1194.207	12.0363	-1.12	3	1673784	-1.88	(free SUMO2)		

Table 3.2 Table of isopeptides identified by ICID analysis in a RanGAP1 SUMOylation reaction

Isopeptides were identified as belonging to RanGAP1 and SUMO2 based on accurate mass assignment to theoretical tryptic peptides. Note that each peptide is supported by either multiple charge states (z) or multiple UBL variant peptides (missed cleaved SUMO2 has an isotopic mass shift of 18 Da rather than 12 Da). A total of 200 hits were tabulated but those without peptide assignments have been truncated, except for the expected free SUMO2 C-terminal peptide. The data table is essentially representative of the ICID output, although some columns have been omitted for presentation clarity.

To confirm that isopeptide assignments are not a result of incorrect feature detection, MS1 spectra were manually validated for correct assignment of monoisotopic peaks. Figure 3.19 presents MS1 spectra identified as SUMO or RanGAP1 isopeptides, all of which were validated as correct. The full ICID results output included a total of 200 isotope coded features, the majority of which did not result in peptide assignments. A repeat ICID analysis on the RanGAP1 SUMOylation reaction with light SUMO2 returned only 5 false positives indicating that almost all of the 200 detected features are a result of isotope labelling.

To demonstrate incorporation with the SUMmOn search engine, SUMmOn searches were conducted for 8 C-terminal SUMO2 peptides (permutations of missed-cleaved, methionine oxidised, and heavy labelled) and compiled with ICID results into a single analysis. 437 SUMmOn assigned spectra passed high resolution precursor filtering which were then filtered to only 90 spectra with ICID support thus significantly minimising manual validation requirements. Only SUMO conjugation sites with multiple instances were considered resulting in rapid validation of RanGAP1 and SUMO2 sites (Table 3.3). However, no spectral support existed in the SUMmOn searches for the SUMO remnant or substrate peptide for the second RanGAP1 K452 site. This modification therefore remains an identification solely based on accurate mass assignment from isotope coding. SUMmOn assigned spectra for the SUMO2 K11 and RanGAP1 K524 isopeptides are presented in Figures 3.20 and 3.21, respectively.

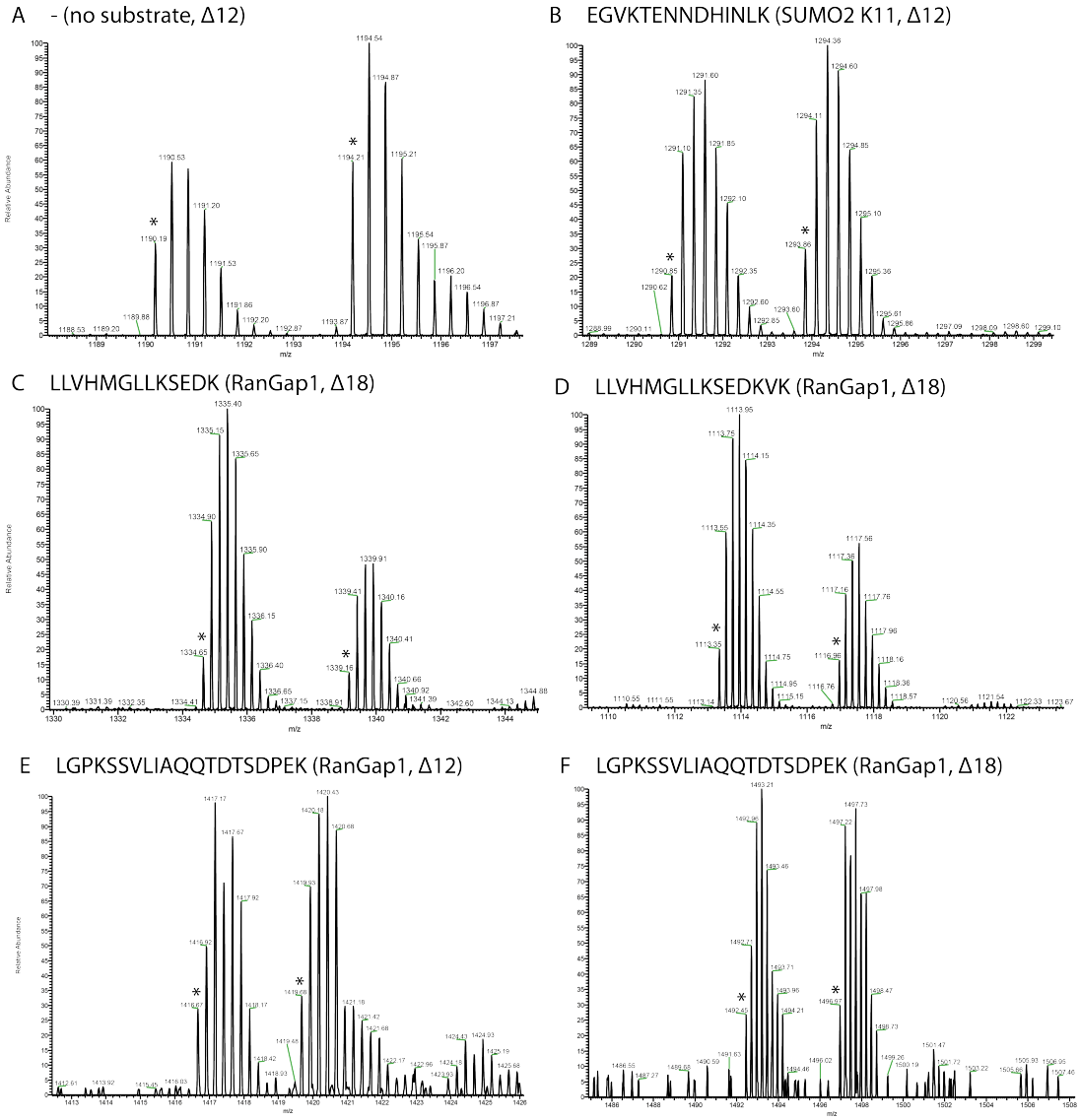


Figure 3.19 Manual validation of MS1 spectra confirms correct detection of isotope coded features in a RanGAP1 *in vitro* reaction

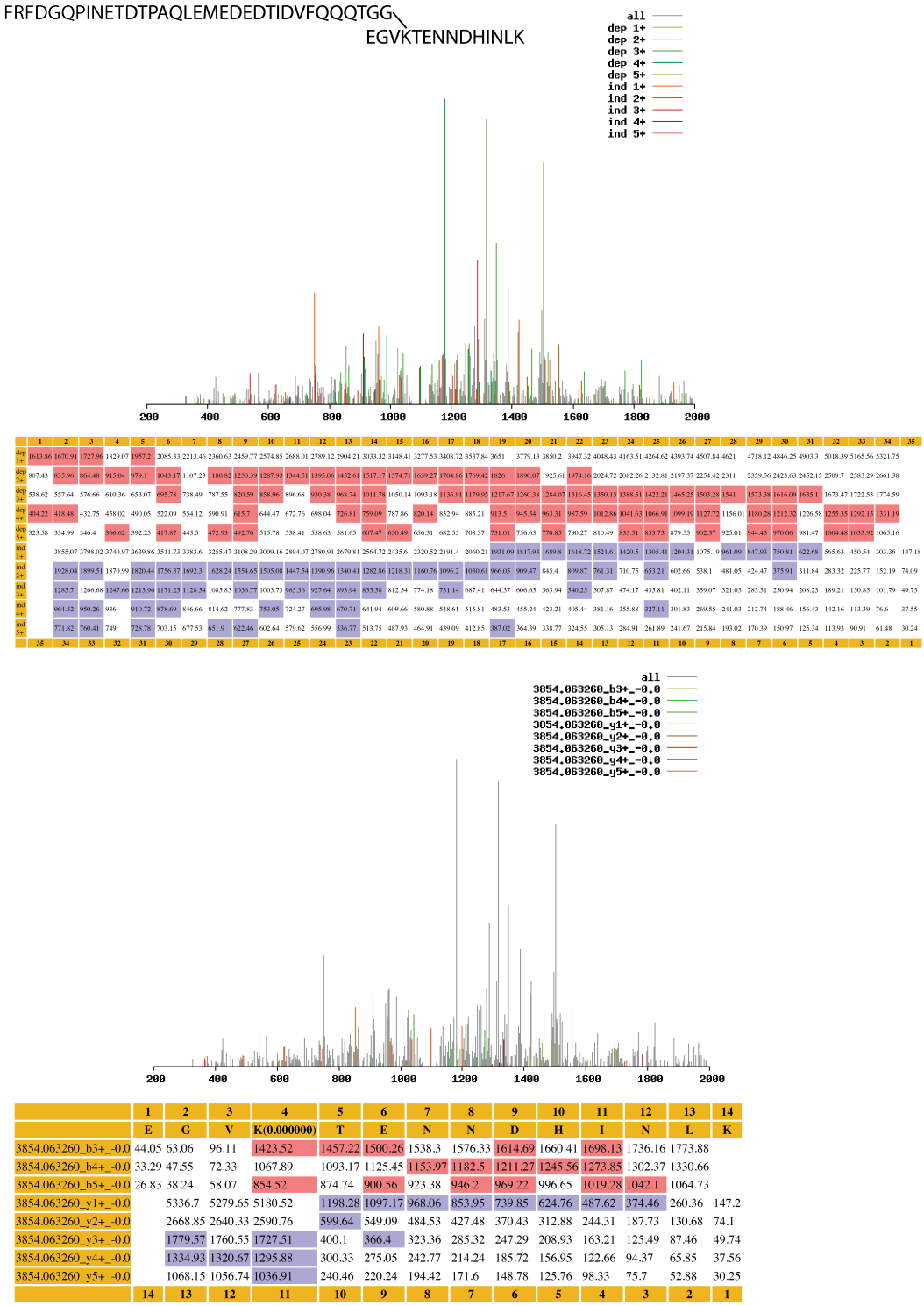
A selection of isotope coded peptide MS1 spectra are presented that were reported as being isotope coded by an automated ICID analysis of a RanGAP1 SUMOylation reaction. Asterisks indicate the reported monoisotope and isotope coded mass interval, which have been correctly identified in all cases for the unmodified C-terminal SUMO2 peptide (A), K11 SUMO2-SUMO2 isopeptide (B), and a selection of RanGAP1 isopeptides (C-F). Δ12 and Δ18 refer to the fully tryptic and missed cleaved SUMO2 C-terminal peptide containing 12 and 18 13C carbon isotopes, respectively.

m/z	isotope	charge	UBL state	modScore (ions)	cndScore (ions)	candidate MWcalc	ppm Error	ICID spectrum (L/H)	pepSeq	description
1093.70911	i=1	z=5	SUMO2_1mc	mod=1.0 (86)	cand=0.95 (17)	1609.806	-1.24	12014/ -	EGVK*TENNDHINLK	SUMO2
1096.71228	i=0	z=5	SUMO2_Mox_1mc	mod=1.0 (81)	cand=0.98 (17)	1609.806	3.19	11871/ -	EGVK*TENNDHINLK	SUMO2
1097.12561	i=0	z=5	SUMO2_1mc_hF	mod=1.0 (97)	cand=0.95 (15)	1609.806	3.43	- /12014	EGVK*TENNDHINLK	SUMO2
1100.32214	i=0	z=5	SUMO2_Mox_1mc_hF	mod=1.0 (91)	cand=1.03 (16)	1609.806	1.18	- /11871	EGVK*TENNDHINLK	SUMO2
1290.84692	i=0	z=4	SUMO2	mod=1.0 (49)	cand=1.01 (12)	1609.806	3	11794/ -	EGVK*TENNDHINLK	SUMO2
1293.85559	i=0	z=4	SUMO2_hF	mod=0.99 (48)	cand=0.96 (11)	1609.806	1.93	- /11794	EGVK*TENNDHINLK	SUMO2
1294.84167	i=0	z=4	SUMO2_Mox	mod=1.0 (54)	cand=1.24 (20)	1609.806	-0.06	11607/ -	EGVK*TENNDHINLK	SUMO2
1297.85181	i=0	z=4	SUMO2_Mox_hF	mod=1.0 (55)	cand=1.31 (17)	1609.806	-0.02	- /11640	EGVK*TENNDHINLK	SUMO2
1366.63513	i=0	z=4	SUMO2_1mc	mod=0.98 (30)	cand=0.95 (5)	1609.806	-0.22	12025/ -	EGVK*TENNDHINLK	SUMO2
1370.63501	i=0	z=4	SUMO2_Mox_1mc	mod=0.97 (38)	cand=0.98 (10)	1609.806	0.62	11871/ -	EGVK*TENNDHINLK	SUMO2
1371.1571	i=0	z=4	SUMO2_1mc_hF	mod=1.0 (45)	cand=0.98 (9)	1609.806	4.81	- /12014	EGVK*TENNDHINLK	SUMO2
1375.15039	i=0	z=4	SUMO2_Mox_1mc_hF	mod=1.0 (34)	cand=1.07 (6)	1609.806	0.85	- /11871	EGVK*TENNDHINLK	SUMO2
1334.64282	i=0	z=4	SUMO2_1mc	mod=1.0 (32)	cand=1.18 (8)	1481.8275	1.52	13807/ -	LLVHMGLLK*SEDK	RanGAP1
1007.28143	i=0	z=5	SUMO2	mod=0.96 (24)	cand=0.96 (5)	1481.8275	1.23	13796/ -	LLVHMGLLK*SEDK	RanGAP1
1339.40088	i=1	z=4	SUMO2_1mc_hF	mod=0.9 (20)	cand=1.55 (5)	1481.8275	-4.36	- /13807	LLVHMGLLK*SEDK	RanGAP1
927.95721	i=0	z=6	SUMO2_1mc	mod=0.97 (23)	cand=0.97 (8)	1708.991	0.36	13279/ -	LLVHMGLLK*SEDKVK	RanGAP1
930.79199	i=1	z=6	SUMO2_Mox_1mc	mod=0.98 (19)	cand=1.12 (5)	1708.991	2.23	13257/ -	LLVHMGLLK*SEDKVK	RanGAP1
1052.7179	i=0	z=5	SUMO2	mod=1.0 (40)	cand=0.93 (11)	1708.991	4.75	13279/ -	LLVHMGLLK*SEDKVK	RanGAP1
1055.91467	i=0	z=5	SUMO2_Mox	mod=1.0 (57)	cand=0.93 (16)	1708.991	2.65	13213/ -	LLVHMGLLK*SEDKVK	RanGAP1
1058.31909	i=0	z=5	SUMO2_Mox_hF	mod=1.0 (25)	cand=0.93 (8)	1708.991	-0.77	- /13213	LLVHMGLLK*SEDKVK	RanGAP1
1113.34949	i=0	z=5	SUMO2_1mc	mod=1.0 (71)	cand=0.93 (9)	1708.991	2.43	13345/ -	LLVHMGLLK*SEDKVK	RanGAP1
1116.54639	i=0	z=5	SUMO2_Mox_1mc	mod=1.0 (73)	cand=0.97 (17)	1708.991	0.54	13224/ -	LLVHMGLLK*SEDKVK	RanGAP1
1120.16296	i=0	z=5	SUMO2_Mox_1mc_hF	mod=1.0 (81)	cand=0.97 (17)	1708.991	4.57	- /12377	LLVHMGLLK*SEDKVK	RanGAP1
1391.9292	i=2	z=4	SUMO2_1mc	mod=0.97 (28)	cand=0.93 (10)	1708.991	-2.98	13345/ -	LLVHMGLLK*SEDKVK	RanGAP1
1395.42737	i=0	z=4	SUMO2_Mox_1mc	mod=0.98 (32)	cand=1.49 (12)	1708.991	-2.17	13235/ -	LLVHMGLLK*SEDKVK	RanGAP1
1400.19763	i=1	z=4	SUMO2_Mox_1mc_hF	mod=1.0 (21)	cand=0.93 (6)	1708.991	0.94	- /13235	LLVHMGLLK*SEDKVK	RanGAP1

Table 3.3 Isopeptides identified by ICID are also independently identified by SUMmOn

An independent MS2 spectral analysis of SUMO2 modified substrates by the SUMmOn search engine, followed by compilation of SUMmOn and ICID results, reveal reciprocal conformation of expected RanGAP1 and SUMO2 isopetpides. Data is presented from the compilation tool which filters SUMmOn searches by accurate precursor mass and requiring supporting evidence from ICID analysis. The data table is essentially representative of the ICID output, although some columns have been omitted for presentation clarity.

Scan	Charge	relPos	modScore	precursor mass	matched ions	TIC	%TIC	BPI
12191	5	1	1.00	1610.8026	86	371464.00	0.45	11885.70

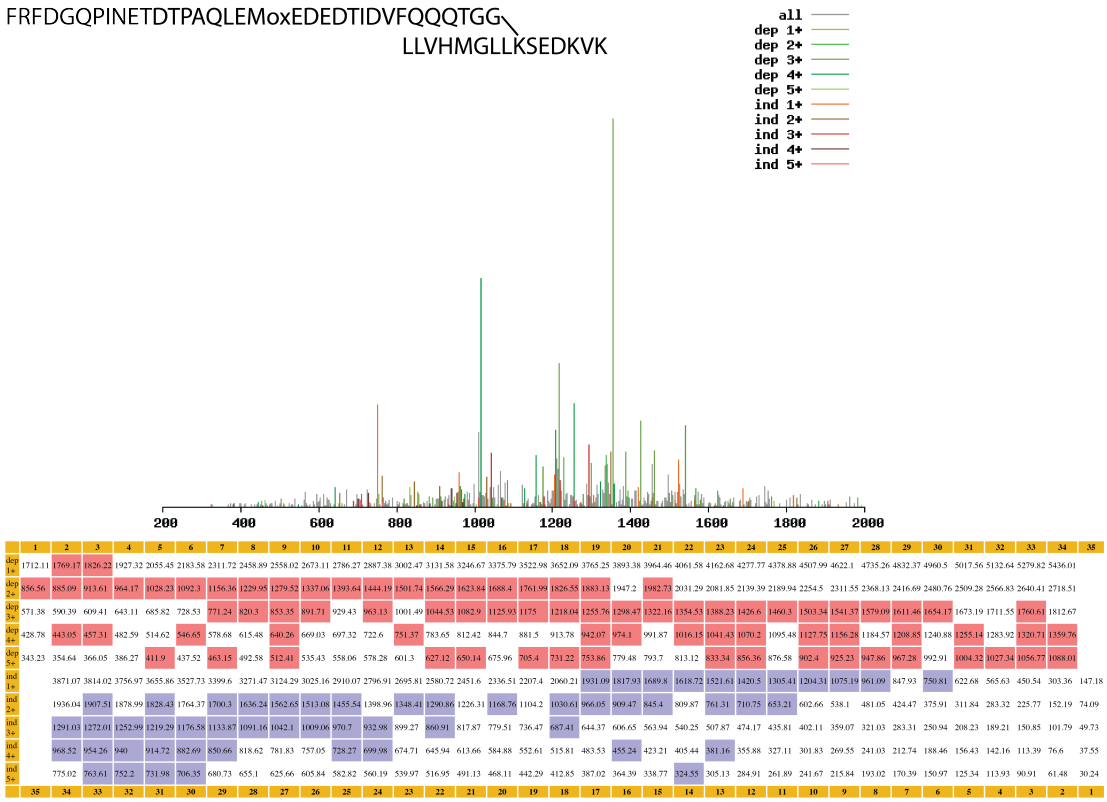


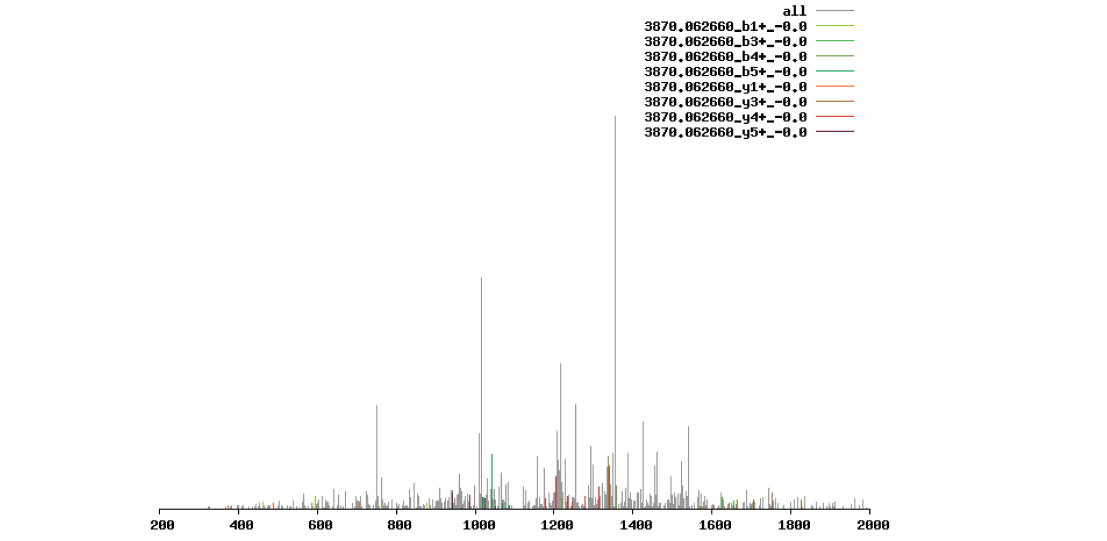
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	E	G	V	K(0.000000)	T	E	N	N	D	H	I	N	L	K
3854.063260_b3+_-0.0	44.05	63.06	96.11	1423.52	1457.22	1500.26	1538.3	1576.33	1614.69	1660.41	1698.13	1736.16	1773.88	
3854.063260_b4+_-0.0	33.29	47.55	72.33	1067.89	1093.17	1125.45	1153.97	1182.5	1211.27	1245.56	1273.85	1302.37	1330.66	
3854.063260_b5+_-0.0	26.83	38.24	58.07	854.52	874.74	900.56	923.38	946.2	969.22	996.65	1019.28	1042.1	1064.73	
3854.063260_y1+_-0.0		5336.7	5279.65	5180.52	1198.28	1097.17	968.06	853.95	739.85	624.76	487.62	374.46	260.36	147.2
3854.063260_y2+_-0.0		2668.85	2640.33	2590.76	599.64	549.09	484.53	427.48	370.43	312.88	244.31	187.73	130.68	74.1
3854.063260_y3+_-0.0		1779.57	1760.55	1727.51	400.1	366.4	323.36	285.32	247.29	208.93	163.21	125.49	87.46	49.74
3854.063260_y4+_-0.0		1334.93	1320.67	1295.88	300.33	275.05	242.77	214.24	185.72	156.95	122.66	94.37	65.85	37.56
3854.063260_y5+_-0.0		1068.15	1056.74	1036.91	240.46	220.24	194.42	171.6	148.78	125.76	98.33	75.7	52.88	30.25
	14	13	12	11	10	9	8	7	6	5	4	3	2	1

Figure 3.20 Spectral support for the SUMO2 K11 isopeptide

Spectra assigned by SUMmOn are presented for the SUMO2-SUMO2 K11 polymerisation site. Fragment ions are independently assigned to the SUMO2 remnant (top) and the substrate peptide (bottom).

Scan	Charge	relPos	modScore	precursor mass	matched ions	TIC	%TIC	BPI
13148	5	1	1.00	1708.9940	73	252539.00	0.43	15615.30





	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	L	L	V	H	M	G	L	L	K(0.000000)	S	E	D	K	V	K
3870.062660_b1+_-0.0	114.17	227.33	326.46	463.6	594.79	651.84	765	878.16	4876.4	4963.48	5092.59	5207.68	5335.86	5434.99	
3870.062660_b3+_-0.0	38.73	76.45	109.49	155.21	198.94	217.95	255.67	293.39	1626.14	1655.16	1698.2	1736.57	1779.29	1812.34	
3870.062660_b4+_-0.0	29.3	57.59	82.37	116.66	149.45	163.72	192.01	220.3	1219.86	1241.63	1273.9	1302.68	1334.72	1359.5	
3870.062660_b5+_-0.0	23.64	46.27	66.1	93.53	119.76	131.18	153.81	176.44	976.09	993.5	1019.33	1042.34	1067.98	1087.8	
3870.062660_y1+_-0.0		5468.02	5354.86	5255.73	5118.59	4987.39	4930.34	4817.18	4704.02	705.79	618.71	489.59	374.5	246.33	147.2
3870.062660_y3+_-0.0		1823.35	1785.63	1752.58	1706.87	1663.14	1644.12	1606.4	1568.68	235.93	206.91	163.87	125.51	82.78	49.74
3870.062660_y4+_-0.0		1367.76	1339.47	1314.69	1280.4	1247.6	1233.34	1205.05	1176.76	177.2	155.43	123.15	94.38	62.34	37.56
3870.062660_y5+_-0.0		1094.41	1071.78	1051.95	1024.52	998.29	986.87	964.24	941.61	141.96	124.55	98.72	75.71	50.07	30.25
	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1

Figure 3.21 Spectral support for the RanGAP1 K542 isopeptide

Spectra assigned by SUMmOn are presented for the SUMO2-RanGAP1 K542. Fragment ions are independently assigned to the SUMO2 remnant (top) and the substrate peptide (bottom).

Additional peptides were found that covered the RanGAP1 K542 SUMO conjugation site, including an oxidised methionine and an additional missed cleavage (Table 3.4). A K271 SAE2 isopeptide was also discovered, which also had an oxidised methionine, and is a known UBC9-mediated SUMOylation site (Truong et al., 2012). These peptides were not originally assigned during the ICID analysis using accurate mass due to strict *in silico* digest parameters. This exemplifies the dependence ICID peptide assignment has on prior sequence assumptions, however, detection of unassigned isotope coded features still provides an effective means of validating isopeptides identified by SUMmOn. Manual validation of features for the SUMmOn identified isopeptides confirmed correct detection of isotope coded monoisotopes (Figure 3.22).

m/z	isotope	charge	UBL state	modScore (ions)	cndScore (ions)	candidate MWcalc	ppm Error	ICID spectrum (L/H)	pepSeq	description
1055.91467	i=0	z=5	SUMO2_s2	mod=1.0 (51)	cand=0.93 (15)	1724.9904	1.80	13213/ -	LLVHMoxGLLK*SEDKVK	RanGAP1
930.79199	i=1	z=6	SUMO2_1mc	mod=0.81 (16)	cand=0.97 (3)	1724.9904	1.42	13257/ -	LLVHMoxGLLK*SEDKVK	RanGAP1
1395.68481	i=1	z=4	SUMO2_1mc	mod=1.0 (43)	cand=0.93 (8)	1724.9904	1.75	13235/ -	LLVHMoxGLLK*SEDKVK	RanGAP1
1116.54639	i=0	z=5	SUMO2_1mc	mod=1.0 (61)	cand=0.93 (13)	1724.9904	-0.25	13224/ -	LLVHMoxGLLK*SEDKVK	RanGAP1
1400.19604	i=1	z=4	SUMO2_1mc_hF	mod=0.99 (31)	cand=0.97 (4)	1724.9904	-0.99	- /13235	LLVHMoxGLLK*SEDKVK	RanGAP1
1120.16296	i=0	z=5	SUMO2_1mc_hF	mod=1.0 (81)	cand=0.97 (15)	1724.9904	3.79	- /12377	LLVHMoxGLLK*SEDKVK	RanGAP1
1387.8728	i=2	z=5	SUMO2	mod=0.9 (35)	cand=0.99 (25)	3382.7708	1.87	16810/ -	AFNSSSFNSNTFLTRLLVHMGLLK*SEDKVK	RanGAP1
1334.64282	i=0	z=4	SUMO2_1mc	mod=1.0 (32)	cand=0.96 (5)	1481.8109	4.63	13807/ -	YLLTMoxDK*LWRK	SAE2
1007.28143	i=0	z=5	SUMO2	mod=0.96 (24)	cand=0.97 (6)	1481.8109	4.53	13796/ -	YLLTMoxDK*LWRK	SAE2
1339.40088	i=1	z=4	SUMO2_1mc_hF	mod=0.9 (20)	cand=0.92 (2)	1481.8109	-1.26	- /13807	YLLTMoxDK*LWRK	SAE2
1334.64172	i=0	z=4	SUMO2_Mox_1mc	mod=0.81 (16)	cand=0.96 (3)	1465.8115	4.65	13807/ -	YLLTMDK*LWRK	SAE2

Table 3.4 Isopeptides identified by SUMmOn are supported by isotope-coded features

Isopeptides uniquely identified by SUMmOn include modifications and miss-cleavages not originally assigned by the ICID analysis due to a strict search space. The unique peptides (or peptide modification state) are supported by detection of isotope coded features.

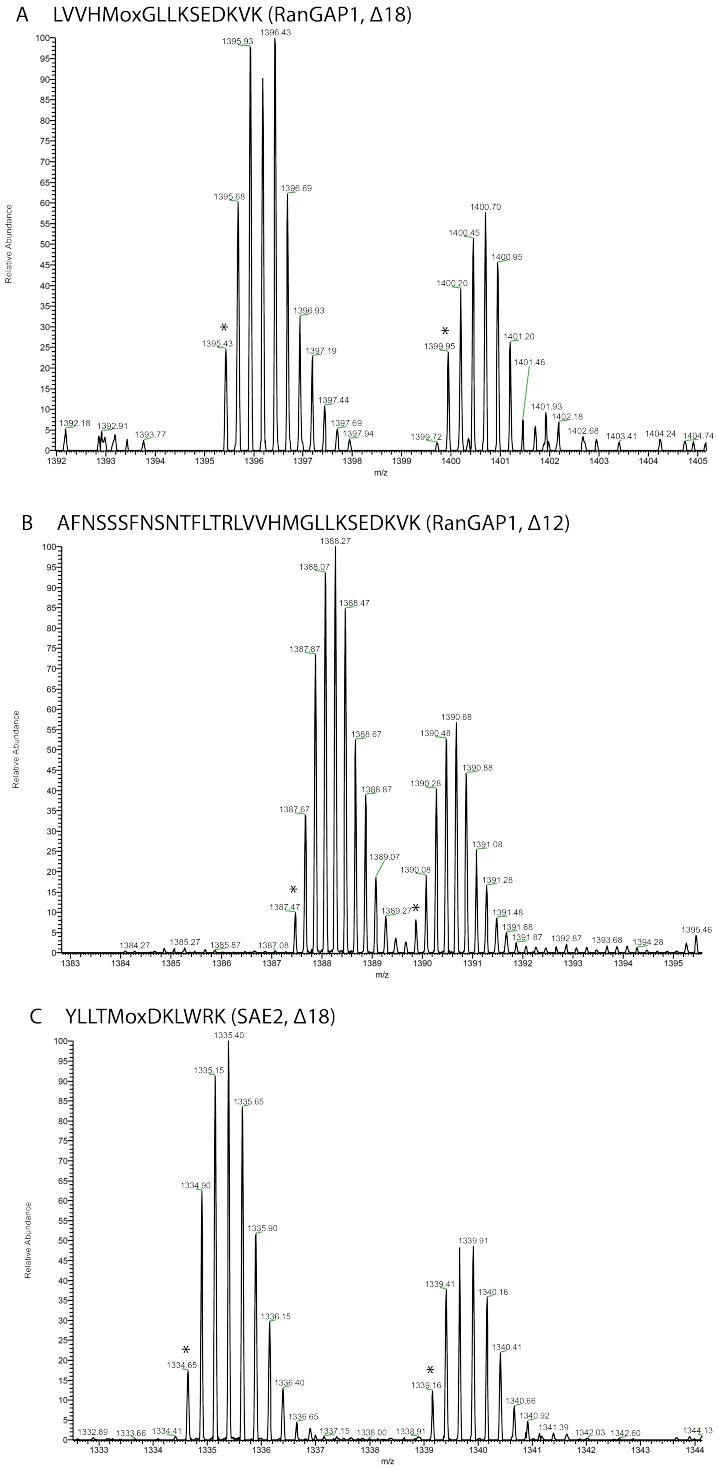


Figure 3.22 Manual validation of MS1 spectra corroborates SUMmOn isotopeptide identifications in a RanGAP1 *in vitro* reaction

Isotope coded peptide MS1 spectra are presented for peptides that were not initially identified in the ICID analysis of a RanGAP1 SUMOylation reaction, but were identified by SUMmOn. Asterisk labels indicate the monoiostope reported by the ICID analysis, which are correct in all cases for the K524 RanGAP1 methionine-oxidised (A) and doubly miss-cleaved (B) isopeptides and a methionine-oxidised SAE2 K271 isopeptide. Δ12 and Δ18 refer to the fully tryptic and missed cleaved SUMO2 C-terminal peptide containing 12 and 18 13C carbon isotopes, respectively.

The ICID module is a valuable addition to SUMmOn for validating identifications. Table 3.5 presents a set of spectra that were all assigned by SUMmOn to another novel RanGAP1 SUMOylation site. Spectra passed the high resolution precursor mass filter, however, inspection of their precursor features (figure 3.23) show that they lack isotopic partners. The compiled data lacked support from ICID features and was correctly filtered out of the dataset. Using concordance between SUMmOn and ICID data is therefore an effective and convenient way to discern between true isopeptides and false positives with minimal interaction. A small high quality dataset can then be given attention for manual validation and confirmation of UBL conjugation sites.

m/z	isotope	charge	UBL state	modScore (ions)	cndScore (ions)	candidate MWCalc	ppm Error	ICID spectrum (L/H)	pepSeq
1105.01343	i=2	z=6	SUMO2_Mox_1mc	mod=0.96 (28)	cand=1.42 (17)	2753.3772	-7.31665	- / -	VSSVFK*DEATVRMoxAVQDAVDALMQK
1105.01318	**i=2**	**z=6**	**SUMO2_Mox_1mc**	**mod=1.0 (38)**	**cand=1.2 (20)**	**2753.3772**	**-7.52781**	**- / -**	**VSSVFK*DEATVRMAVQDAVDALMoxQK**
1054.16125	**i=0**	**z=6**	**SUMO2_Mox**	**mod=0.95 (33)**	**cand=1.32 (24)**	**2753.3772**	**2.30831**	**- / -**	**VSSVFK*DEATVRMoxAVQDAVDALMQK**
1325.83447	i=0	z=5	SUMO2_1mc_hF	mod=0.89 (26)	cand=0.96 (11)	2753.3772	-1.22187	- / -	VSSVFK*DEATVRMAVQDAVDALMoxQK
1319.02673	i=0	z=5	SUMO2_1mc	mod=0.84 (65)	cand=0.96 (27)	2737.3779	1.92566	- / -	VSSVFK*DEATVRMAVQDAVDALMQK
947.15662	i=1	z=7	SUMO2_Mox_1mc	mod=0.97 (37)	cand=0.96 (17)	2753.3772	-5.51349	- / -	VSSVFK*DEATVRMAVQDAVDALMoxQK
1102.02063	i=0	z=6	SUMO2_1mc	mod=0.9 (59)	cand=1.19 (27)	2753.3772	-0.55958	- / -	VSSVFK*DEATVRMoxAVQDAVDALMQK
1322.427	i=1	z=5	SUMO2_1mc	mod=0.8 (36)	cand=1.54 (30)	2753.3772	1.74603	- / -	VSSVFK*DEATVRMAVQDAVDALMoxQK

Table 3.5 False positive SUMmOn assignments lack support from ICID features

A potentially novel RanGAP1 isopeptide was assigned by SUMmOn to multiple peptide variants. Precursor spectra are presented in Figure 3.23 for those in bold.

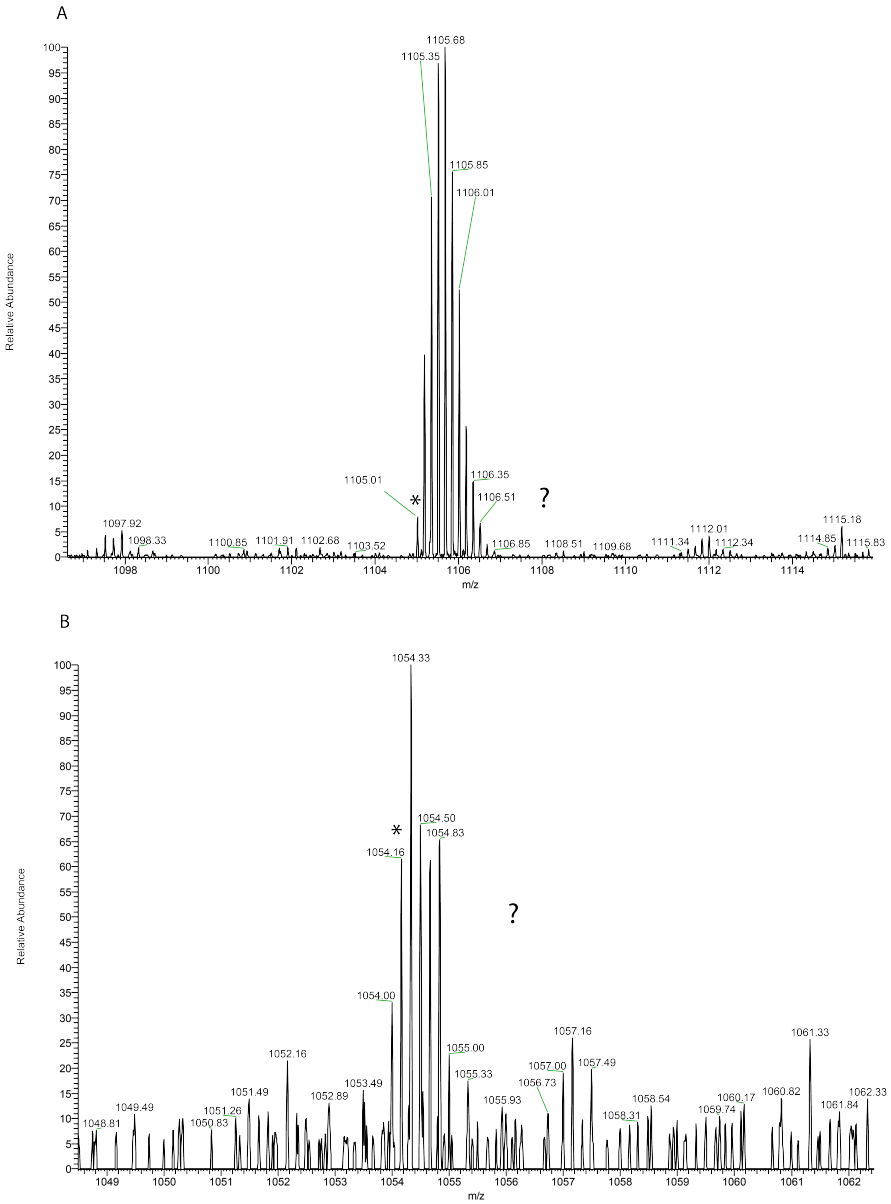


Figure 3.23 False positive SUMmOn assignments to a RanGAP1 isopepide lack isotopic partners

Manual validation of precursors to potential RanGAP1 SUMOylation sites (bold entries in Table 3.5) are indicated (*) and lack isotopic partners at the expected mass interval (?). This confirms that the SUMmOn assignment is a false positive and that lack of support from the ICID analysis has correctly contradicted the MS2 spectral interpretation.

3.3.6 Application to uncharacterised substrate - RNMT

RNA guanine-7 methyltransferase (RNMT) methylates the N-7 position during formation of the 5' mRNA cap which is essential for mRNA translation (Cowling, 2010). RNMT also requires interaction with RNMT-Activating Mini protein (RAM) for efficient cap methylation (Gonatopoulos-Pournatzis et al., 2011). In unpublished

work in the lab of Dr Cowling (Dundee University), HeLa cells transiently transfected with expression vectors encoding RNMT, RNMT was detected as a ladder suggesting that RNMT was modified by a UBL. The ladder was increased by MG132 suggesting that the modification leads to proteasomal degradation. Curiously, when co-transfected with RAM, the ladder disappeared suggesting that the RAM-RNMT interaction stabilised RNMT. SUMO modification was hypothesised, although the nature of the modification *in vivo* has not been confirmed. Investigation into *in vitro* SUMOylation revealed that RNMT also results in a modification ladder under *in vitro* conditions (Figure 3.24). It is not clear from the gel image whether the RNMT SUMOylation ladder is a result of multiple mono-SUMOylations or formation of SUMO chains on RNMT.

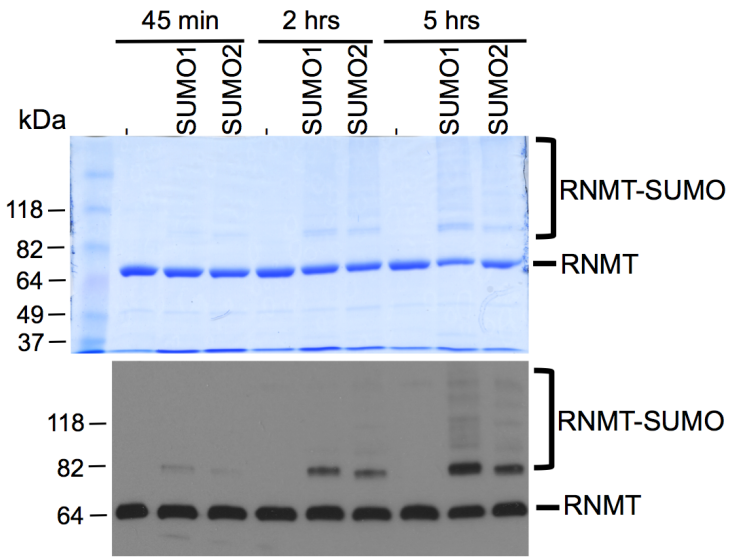


Figure 3.24 RNMT is SUMOylated with either SUMO1 or SUMO2 by UBC9 *in vitro*

Recombinant monomeric RNMT was subjected to *in vitro* sumoylation in the presence of SUMO1, SUMO2 or no SUMO protein (-) for the time points indicated above the panels. The reaction products were analysed by Coomassie stained SDS-polyacrylamide gel (upper panel) or western blot to detect RNMT (lower panel). RNMT and sumoylated RNMT (RNMT-SUMO) are indicated. Figure reproduced with permission from Thomas Gonatopoulos Pournatzis (2012) PhD thesis.

The RNMT (isoform 1) SUMOylation was repeated using a mixture of light and isotopically heavy SUMO2. LCMS data from the tryptic digest was subjected to ICID analysis and SUMmOn analysis followed by SUMmOn/ICID data compilation. Manual validation required multiple instances of a lysine modification site and with support from ICID features (exactly as done for the RanGAP1 validation experiment). Final results were manually validated to confirm that precursor monoisotopes were correctly identified. 17 modification sites were detected (Table 3.6), although only 16 could reliably be assigned to RNMT.

A surprisingly large selection of lysines were found to be modified indicating a very low degree of specificity. Long reactions times are also required to achieve a significant level of modification. This would suggest that RNMT is not a specific target under the current conditions and *in vivo* modification is mediated by an additional factors, if the modification observed *in vivo* is indeed SUMOylation. Given the lack of consensus site for UBC9 mediated modification, it would seem likely that an E3 ligase is required to increase specificity and efficiency. The modification area over the protein is large (Figure 3.25) and spans both the N-terminal domain (1-120) and the catalytic domain (121-476). The non-catalytic N-terminal domain, which is required for recruitment to transcription initiation sites, does not bind RAM (Aregger and Cowling, 2013). It would seem unlikely that RAM binding would protect from SUMOylation at these sites suggesting that the modifications in the N-terminal domain are non-specific reactivity.

m/z	isotope	charge	UBL state	modScore (ions)	cndScore (ions)	candidate MWcalc	ppm Error	ICID spectrum (L/H)	pepSeq	lysine#
1376.63794	i=0	z=4	SUMO2_1mc	mod=1.0 (33)	cand=0.94 (8)	1649.7897	4.776129	27632/ -	GPLGSANSAK*AEEYEK	K6
1101.51074	i=0	z=5	SUMO2_1mc	mod=0.85 (76)	cand=0.94 (26)	1649.7897	3.831102	27294/ -	GPLGSANSAK*AEEYEK	
787.07574	i=0	z=7	SUMO2_1mc	mod=0.98 (112)	cand=0.94 (25)	1649.7897	-3.08556	26657/ -	GPLGSANSAK*AEEYEK	
1381.39709	i=1	z=4	SUMO2_1mc_hF	mod=0.97 (31)	cand=0.94 (13)	1649.7897	-0.11854	- /27645	GPLGSANSAK*AEEYEK	
1380.88342	i=1	z=4	SUMO2_Mox_1mc	mod=0.99 (28)	cand=0.94 (6)	1649.7897	1.818582	27242/ -	GPLGSANSAK*AEEYEK	
1386.14148	i=4	z=4	SUMO2_Mox_1mc_hF	mod=0.98 (41)	cand=0.94 (12)	1649.7897	-5.07163	- /27229	GPLGSANSAK*AEEYEK	
1109.10938	i=4	z=5	SUMO2_Mox_1mc_hF	mod=1.0 (72)	cand=0.94 (16)	1649.7897	-9.81328	- /26813	GPLGSANSAK*AEEYEK	
1277.07434	i=0	z=4	SUMO2	mod=1.0 (40)	cand=0.92 (9)	1554.7235	1.507352	28152/ -	AEEYEK*MSLEQAK	K12
1353.11438	i=1	z=4	SUMO2_1mc	mod=1.0 (19)	cand=0.92 (9)	1554.7235	-0.93396	28503/ -	AEEYEK*MSLEQAK	
1082.49329	i=0	z=5	SUMO2_1mc	mod=1.0 (55)	cand=0.92 (14)	1554.7235	0	28568/ -	AEEYEK*MSLEQAK	
1357.63354	i=1	z=4	SUMO2_1mc_hF	mod=1.0 (31)	cand=0.92 (4)	1554.7235	2.089113	- /28503	AEEYEK*MSLEQAK	
1086.10913	i=0	z=5	SUMO2_1mc_hF	mod=0.99 (62)	cand=0.92 (17)	1554.7235	3.480313	- /28568	AEEYEK*MSLEQAK	
1280.33374	i=1	z=4	SUMO2_hF	mod=1.0 (43)	cand=0.92 (11)	1554.7235	0.340731	- /28152	AEEYEK*MSLEQAK	
1132.77246	i=0	z=4	SUMO2_1mc	mod=1.0 (69)	cand=0.92 (3)	674.3599	-1.28005	26709/ -	DTPSK*K	K80
1137.53931	i=1	z=4	SUMO2_1mc_hF	mod=1.0 (83)	cand=1.02 (7)	674.3599	-0.45163	- /26709	DTPSK*K	
1212.56726	i=1	z=4	SUMO2	mod=1.0 (45)	cand=0.91 (11)	1295.6973	0.462861	29127/ -	K*LDPEIVPEEK	K83
1288.35779	i=1	z=4	SUMO2_1mc	mod=1.0 (76)	cand=0.91 (15)	1295.6973	-1.01971	29374/ -	K*LDPEIVPEEK	
1030.68982	i=0	z=5	SUMO2_1mc	mod=0.96 (42)	cand=0.91 (16)	1295.6973	1.726999	29335/ -	K*LDPEIVPEEK	
1292.62585	i=0	z=4	SUMO2_1mc_hF	mod=1.0 (69)	cand=0.91 (14)	1295.6973	1.953388	- /29374	K*LDPEIVPEEK	
1215.57788	i=1	z=4	SUMO2_hF	mod=0.99 (63)	cand=0.91 (16)	1295.6973	0.934741	- /29127	K*LDPEIVPEEK	
1173.80322	i=1	z=4	SUMO2	mod=1.0 (61)	cand=1.2 (12)	1140.639	0.925411	29062/ -	K*IALEDVPEK	K127
1249.34338	i=0	z=4	SUMO2_1mc	mod=1.0 (72)	cand=1.16 (11)	1140.639	-0.24013	29270/ -	K*IALEDVPEK	
1000.07458	i=2	z=5	SUMO2_1mc	mod=1.0 (69)	cand=0.95 (13)	1140.639	-3.14177	29322/ -	K*IALEDVPEK	
1254.3606	i=2	z=4	SUMO2_1mc_hF	mod=1.0 (69)	cand=1.29 (17)	1140.639	0.13752	- /29283	K*IALEDVPEK	
1003.48724	i=1	z=5	SUMO2_1mc_hF	mod=1.0 (68)	cand=1.13 (11)	1140.639	-1.88443	- /29335	K*IALEDVPEK	
1176.56067	i=0	z=4	SUMO2_hF	mod=1.0 (64)	cand=1.16 (15)	1140.639	-0.55246	- /29075	K*IALEDVPEK	
1253.84705	i=2	z=4	SUMO2_Mox_1mc	mod=1.0 (66)	cand=1.29 (15)	1140.639	2.370704	28971/ -	K*IALEDVPEK	
1258.10437	i=1	z=4	SUMO2_Mox_1mc_hF	mod=1.0 (69)	cand=0.95 (12)	1140.639	-3.1307	- /28971	K*IALEDVPEK	
1281.60889	i=1	z=4	SUMO2_1mc	mod=1.0 (73)	cand=0.91 (14)	1268.6976	-0.2253	28373/ -	IALEDVPEK*QK	K137
1286.12695	i=1	z=4	SUMO2_1mc_hF	mod=0.97 (49)	cand=0.91 (17)	1268.6976	2.108073	- /28373	IALEDVPEK*QK	
1193.3158	i=1	z=4	SUMO2_1mc	mod=0.99 (54)	cand=0.88 (7)	915.5178	1.329279	30960/ -	GGDLLK*WK	K214
1197.57825	i=0	z=4	SUMO2_1mc_hF	mod=1.0 (72)	cand=0.88 (7)	915.5178	-0.167	- /30960	GGDLLK*WK	
1079.50476	i=1	z=4	SUMO2_1mc	mod=1.0 (60)	cand=0.67 (2)	460.2798	0.056739	28204/ -	WK*K	K216
1083.77307	i=0	z=4	SUMO2_1mc_hF	mod=1.0 (52)	cand=0.67 (3)	460.2798	3.806147	- /28204	WK*K	
1304.26294	i=1	z=3	SUMO2	mod=0.99 (49)	cand=0.67 (3)	359.2281	-0.39741	26124/ -	K*GR	K217
1054.24353	i=1	z=4	SUMO2_1mc	mod=1.0 (47)	cand=0.67 (2)	359.2281	1.646915	26878/ -	K*GR	
1058.505	i=0	z=4	SUMO2_1mc_hF	mod=1.0 (64)	cand=0.67 (1)	359.2281	-0.96835	- /26865	K*GR	
1308.27588	i=1	z=3	SUMO2_hF	mod=0.9 (43)	cand=0.67 (1)	359.2281	-0.7529	- /26124	K*GR	
1203.04712	i=1	z=4	SUMO2_1mc	mod=1.0 (74)	cand=0.86 (7)	954.4229	5.516201	27476/ -	YEDMK*NR	K244
1207.3125	i=0	z=4	SUMO2_1mc_hF	mod=1.0 (73)	cand=0.86 (4)	954.4229	6.439923	- /27463	YEDMK*NR	
1179.05066	i=1	z=4	SUMO2_1mc	mod=1.0 (84)	cand=0.86 (9)	858.4559	1.621008	26657/ -	IK*NNENK	K404
1183.56311	i=1	z=4	SUMO2_1mc_hF	mod=1.0 (66)	cand=0.86 (5)	858.4559	-0.60305	- /26644	IK*NNENK	
1403.99133	i=0	z=3	SUMO2	mod=1.0 (28)	cand=0.8 (2)	659.4152	-0.02374	28867/ -	MLLK*R	K413
1053.49377	i=1	z=4	SUMO2	mod=0.98 (40)	cand=0.8 (3)	659.4152	-2.29119	28841/ -	MLLK*R	
1505.38379	i=1	z=3	SUMO2_1mc	mod=1.0 (41)	cand=0.8 (1)	659.4152	0.984245	29192/ -	MLLK*R	
1129.03845	i=0	z=4	SUMO2_1mc	mod=1.0 (71)	cand=0.8 (2)	659.4152	0.642139	29205/ -	MLLK*R	
903.63153	i=1	z=5	SUMO2_1mc	mod=1.0 (67)	cand=0.8 (5)	659.4152	-0.85322	29205/ -	MLLK*R	
907.24329	i=1	z=5	SUMO2_1mc_hF	mod=1.0 (69)	cand=0.8 (5)	659.4152	-1.20254	- /29205	MLLK*R	
1133.28723	i=1	z=4	SUMO2_Mox_1mc	mod=0.81 (52)	cand=0.8 (2)	659.4152	-0.05625	28152/ -	MLLK*R	
1137.55322	i=0	z=4	SUMO2_Mox_1mc_hF	mod=1.0 (61)	cand=0.8 (3)	659.4152	1.494436	- /28152	MLLK*R	
1494.19751	i=0	z=4	SUMO2_1mc	mod=0.94 (27)	cand=3.35 (6)	2120.0459	1.422168	28958/ -	MQALEPYPANESSK*LVSEK	K428
1498.71338	i=0	z=4	SUMO2_1mc_hF	mod=1.0 (26)	cand=2.95 (16)	2120.0459	1.951674	- /28958	MQALEPYPANESSK*LVSEK	
1365.13513	i=1	z=4	SUMO2_1mc	mod=1.0 (36)	cand=1.05 (7)	1602.7889	2.297392	27632/ -	LVSEK*VDDYEHAAK	K433
1092.10791	i=0	z=5	SUMO2_1mc	mod=1.0 (79)	cand=1.07 (12)	1602.7889	1.410117	27671/ -	LVSEK*VDDYEHAAK	
1369.65076	i=1	z=4	SUMO2_1mc_hF	mod=1.0 (37)	cand=1.07 (6)	1602.7889	2.69138	- /27645	LVSEK*VDDYEHAAK	
1292.35095	i=1	z=4	SUMO2_hF	mod=1.0 (59)	cand=1.01 (8)	1602.7889	1.014624	- /27008	LVSEK*VDDYEHAAK	
1065.48535	i=1	z=5	SUMO2_1mc	mod=1.0 (65)	cand=0.92 (13)	1468.6656	2.786524	27658/ -	VDDYEHAAK*YMK	K442
1069.09558	i=1	z=5	SUMO2_1mc_hF	mod=1.0 (82)	cand=0.92 (12)	1468.6656	1.07474	- /27671	VDDYEHAAK*YMK	
1220.31323	i=0	z=4	SUMO2_1mc	mod=1.0 (84)	cand=0.99 (6)	1024.5125	0.962868	27385/ -	YMK*NSQVR	K445
976.65466	i=1	z=5	SUMO2_1mc	mod=0.93 (35)	cand=0.88 (5)	1024.5125	2.978535	27437/ -	YMK*NSQVR	
1225.08008	i=1	z=4	SUMO2_1mc_hF	mod=1.0 (71)	cand=1.04 (8)	1024.5125	1.723357	- /27398	YMK*NSQVR	
980.06683	i=0	z=5	SUMO2_1mc_hF	mod=0.94 (56)	cand=0.88 (6)	1024.5125	3.754846	27450/27450	YMK*NSQVR	
1224.56116	i=1	z=4	SUMO2_Mox_1mc	mod=0.99 (75)	cand=0.98 (9)	1024.5125	-0.37871	26800/ -	YMK*NSQVR	
1229.07483	i=1	z=4	SUMO2_Mox_1mc_hF	mod=0.96 (57)	cand=0.99 (8)	1024.5125	-1.51638	- /26800	YMK*NSQVR	
1284.92505	i=0	z=3	SUMO2	mod=1.0 (37)	cand=0.53 (2)	302.2066	2.490418	26137/ -	K*R	K58,81,104,106,196
1039.73669	i=0	z=4	SUMO2_1mc	mod=1.0 (63)	cand=0.53 (1)	302.2066	1.082005	26865/ -	K*R	ambiguous
1044.5033	i=1	z=4	SUMO2_1mc_hF	mod=1.0 (50)	cand=0.53 (1)	302.2066	1.734078	- /26878	K*R	
1043.7323	i=0	z=4	SUMO2_Mox_1mc	mod=1.0 (62)	cand=0.53 (1)	302.2066	-1.9162	26514/ -	K*R	

Table 3.6 Table of RNMT SUMOylation sites identified by SUMmOn with support from ICID analysis

17 modification sites were identified with high mass accuracy precursors, high SUMmOn scores, and supported from having isotope coded UBL modifications, and observing multiple peptide species. At least one species for each peptide sequences was manually validated to confirm accurate detection of precursor monoisotopes. One SUMO isopeptide with a very short substrate peptide (KR) could be assigned to 5 different location but could also be mapped to SAE2 and therefore cannot be reliably assigned to RNMT.

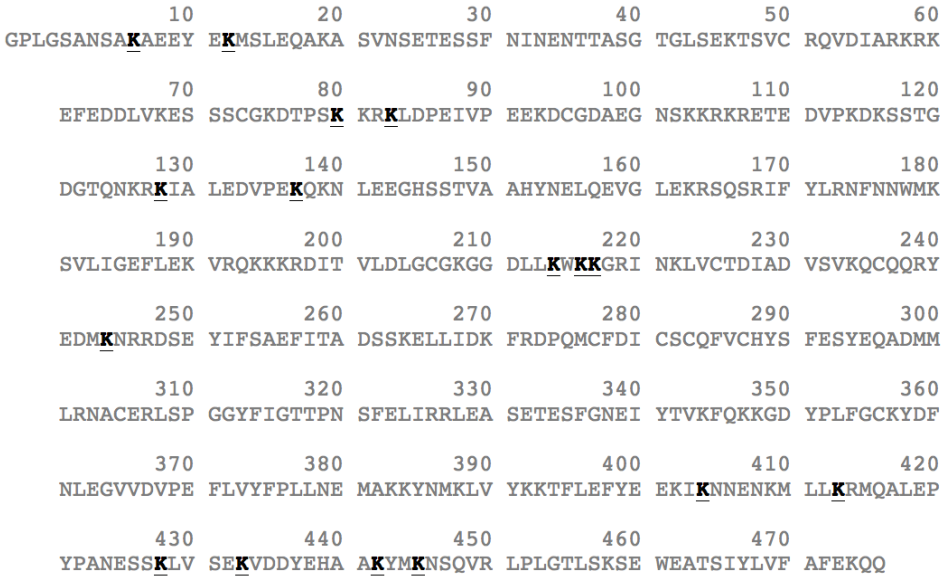


Figure 3.25 A sequence mapping of 16 SUMO modification sites on RNMT

16 SUMOylated lysines are mapped onto the RNMT (isoform 1) sequence (bold/underlined). Amino acid numbering is according to WT human sequence (O43148 MCES_HUMAN isform1) but with an N-terminal extension due to an N-terminal His-tag (the His-tag was cleaved off during purification). Modification sites occur in the N-terminal domain and the catalytic domain (121-476).

To validate if any of these sites are genuine, site directed mutagenesis could be done for the more abundant isopeptide signals. An analysis of lysine mutants may display a phenotype in vivo, or reveal reduced levels of modification by western blot if a major site is perturbed. Better support would come from in vivo detection without mutagenesis. A tandem substrate and SUMO immunoprecipitation would provide

strong support to RNMT being a genuine SUMO substrate. This raises the possibility of direct MS confirmation of in vivo isopeptides now that we have an *in vitro* reference.

3.3.7 Towards *in vivo* isopeptide identification

Mass spectrometry offers powerful techniques for identification and quantification of proteins and their modifications. Application to *in vivo* systems offer the greatest insight into the dynamic nature of the proteome, however, application of these technologies to complex UBL isopeptides *in vivo* is extremely rare. There are significant technical limitations that prohibit efficient analysis of UBL isopeptides and this field of research lags well behind that of simple modifications like phosphorylation and ubiuqitin diglycine remnants. This section on *in vivo* UBL analysis presents future prospects for *in vivo* analysis of complex UBL isopeptides. The main challenges faced will be discussed, and some preliminary results will be presented to support proposed directions for future research.

3.3.7.1 Validation of *in vitro* targets

The biological significance of *in vitro* reactions should always be questioned since the reaction is occurring outside the their natural cellular context. Validating potential modification sites are often approached through mutational analysis. Mutation of lysines can however result in the UBL modification jumping to neighbouring lysines (Klug et al., 2013). A more direct validation on native substrates would have greater biological relevance. *In vitro* identified isopeptides provide a valuable dataset of candidate isopeptides which may occur *in vivo* and can be used as references to target isopeptides in cell and tissue extracts.

The concept of an *in vitro* to *in vivo* strategy was proposed by Matic et al. (2008), where SUMO-SUMO isopeptides were identified from *in vitro* SUMO polymerisation reactions. SUMO chains were then purified from HeLa cells and chains identified using the *in vitro* precursor ion, elution time, and fragmentation spectra as a guide. This strategy was demonstrated only for SUMO polymerisation sites, and was achieved through manual validation against "virtual" linearised peptides. A modified strategy generalised for UBL modified proteins is outlined in figure 3.26.

A major limitation in this strategy is the initial identification of isopeptides which has been a significant challenge even in relatively simple *in vitro* reactions. The methodologies presented in this chapter make the *in vitro* workflow relatively straightforward and the major bottleneck in the strategy has now moved from the *in vitro* to the *in vivo* side. Methods for purifying UBLs like SUMO have recently made headway with SUMO interacting domains (Bruderer et al., 2011) and generation of anti-SUMO antibodies (Becker et al., 2013). Additional enrichment strategies may also be available via substrate specific antibodies. UBL and/or substrate tagging is also feasible for many cell lines and model organisms. Isotope labelled isopeptides from *in vitro* reactions can also be used as spike-in references to aid detection and quantification. The main limitations now lie in MS targeting and spectral interpretation of *in vivo* derived isopeptides.

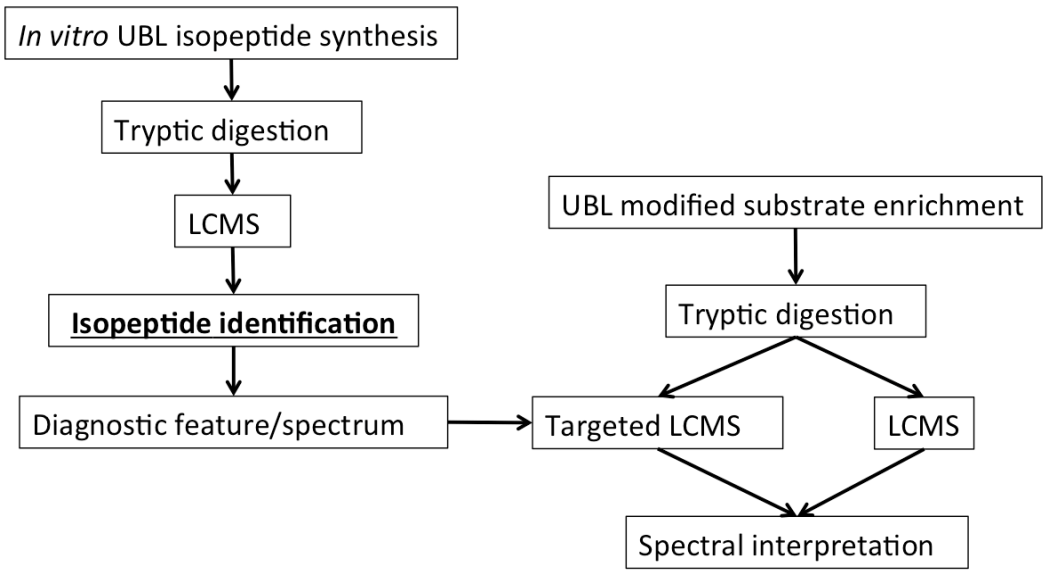


Figure 3.26 Strategy for identifying *in vivo* UBL modification sites from *in vitro* identified UBL isopeptides

Identification of UBL isopeptides from *in vitro* reactions defines spectral and chromatographic properties that can be used to aid *in vivo* detection and validation/invalidation of UBL conjugation sites. Model based on the strategy proposed by Matic et al. (2008). The research in this chapter primarily addresses the challenges for *in vitro* isopeptide identification.

3.3.7.2 Direct detection of *in vivo* isopeptides

Not all UBL modification can be produced *in vitro* due to substrates being difficult to express or the necessary E3 being unknown. Direct detection of isopeptides without *in vitro* guidance would therefore be a valuable capability. The previous strategy used isotopic UBL expressed with heavy amino acids which is only applicable to *in vitro* reactions. An alternative isotope coding strategy for isotope labelling is presented in Fig 3.27. In contrast to the previous method, isotope coding comes from chemical labelling of peptide N-termini. Since isopeptides have two N-termini they will receive two labels. Every peptide gains an isotopic partner but isopeptides can be distinguished from linear peptides by their double spacing. To ensure isotope labelling is specific to N-termini, primary amines are blocked at the protein level prior to digestion. This has the effect of blocking the N-terminus, therefore N-terminal UBL modifications and

modifications on a lysine within the first peptide will not be amenable to analysis by this method. Trypsin will also no longer cleave at lysine making the digest arginine specific, although the method should work equally well with alternative enzymes if necessary. Although the sample processing is more involved than the *in vitro* protocol, the method benefits from being applicable to UBL modified proteins from any source, to protein modification by any of the UBL family, and also simultaneous analysis of mixed UBL modifications. Software requirements to deal with detection of isotope pairs have already been implemented for the *in vitro* protocol and is equally applicable here. The ability to detect and target isopeptides without any sequence assumptions is particularly important in this scenario as isopeptide precursors cannot be predicted in complex samples.

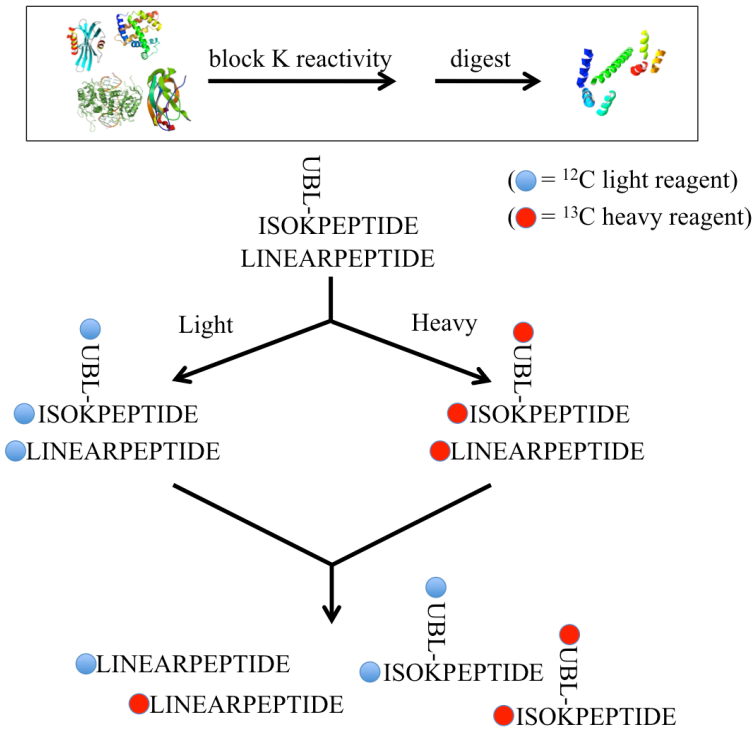


Figure 3.27 A strategy isotope coding isopeptides by chemical modification with isotopic reagents

Amine reactivity is blocked at the protein level prior to proteolytic digestion to expose N-terminal peptides amines, of which there are two of on isopeptides. Divided samples are labelled with light or heavy amine reactive reagents (e.g. mTRAQ) and recombined. Isopeptides can be differentiated from linear peptides by their double isotope spacing.

To demonstrate the viability of the proposed isotopic labelling protocol, the isopeptide discovery process was applied to a mixture of K11/K48/K63 ubiquitin dimers. Ubiquitin amine reactivity was blocked using nicotinic NHS ester, and tryptically exposed N-terminal amines were labelled with either light or +4 isotopically heavy mTRAQ reagents (N-methylpiperazine acetic acid NHS ester). An inclusion list was generated from an initial DDA analysis using the automated ICID software and the sample re-acquired with a targeted DDA analysis (Figure 3.28). In the DDA analysis, 3040 MS2 spectra were acquired accounting for 12 isopeptide spectra. All isopeptides spectra identified belonged to K48 and K63. K11 ubiquitin isopeptides were not identified as a result of blocking the N-terminus and all lysines in the N-terminal half of ubiquitin. During the targeted run, only 233 MS2 spectra were acquired, which included 27 spectra to ubiquitin isopeptides but no linear peptides were identified.

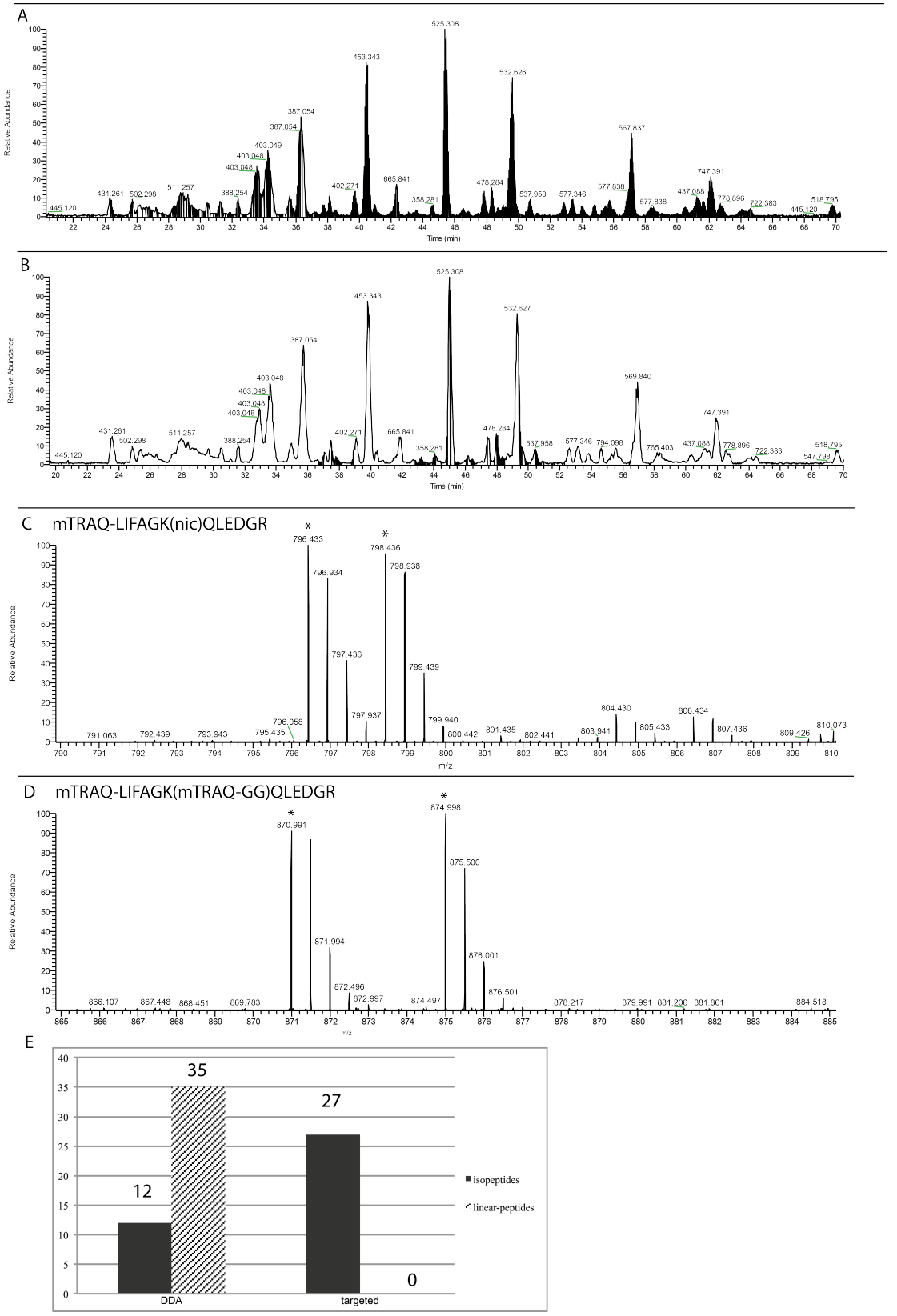


Figure 3.28 Isotopic chemical labelling of ubiquitin dimers enables automated targeted acquisition of ubiquitin isopeptides

Ubiquitin dimers are used as a test case for isotopic chemical labelling as a strategy for targeting isopeptides. Ubiquitin lysines are nicotinic acid (nic) blocked, digested and mTRAQ labelled. An initial DDA analysis (A) is analysed to generate an inclusion list for a targeted MS analysis (B). Shaded regions under the chromatograms indicate

MS2 events. Peptide labels create a 4 Da interval for linear peptides (C) whereas isopeptides are detected at an 8 Da interval (D). Spectral counting of ubiquitin linear peptides and isopeptides (E) show that re-targeting isopeptides resulted in an effective analysis isopeptides without linear peptides interfering with analysis efficiency. This confirms the chemical labelling strategy to be viable option for the study of complex UBL isopeptides for *in vivo* derived samples.

This test case demonstrated that chemical labelling with isotopic reagents can produce signals required for discriminating isopeptides from linear peptides. With this new strategy, isopeptides can be detected and an inclusion list created in an unsupervised manner. A targeted MS analysis can successfully re-acquire isopeptide spectra with a more efficient use of MS cycle time. The proposed new strategy for isotope coding isopeptides offers exciting prospects for broadening the scope of UBL analysis beyond *in vitro* studies. Note however that ubiquitin isopeptides were used to validate the procedure because their spectra are easily interpreted. This does not represent the spectral complexity we would expect from SUMO isopeptides. Nor did the sample offer the complexity and low abundance challenge expected from cellular extracts. Efficient targeting in complex samples requires further attention to move forward in this area of research.

3.3.7.3 MS targeting isopeptides

For the *in vitro* reactions, targeting was not essential for identification of isopeptides as they were sufficiently intense for detection by DDA acquisition. Initial attempts to collect additional spectra on the RNMT sample with targeted LCMS resulted in only a subset of isopeptides being re-acquired. Given that monisotopes from isopeptides were manually validated as correct, this would indicate a technical limitation in the MS targeting method. Targeting *in vivo* isopeptides will be a greater challenge and resolving targeting efficiency will be essential for implementation of the proposed

methodology. Targeting parameters on the Orbitrap Velos software control were investigated for their effect on targeting efficiency. 25 manually validated tryptic peptides were selected from a complex whole yeast tryptic digest at an intensity of 10,000-20,000 counts, which corresponds to ~1/5000 signal relative to base peak. These peptides were identified in an initial DDA analysis, they are low abundance features that are consistently detectable above noise, and they represent a realistic targeting challenge. MS targeted data acquisition was repeated while modifying MS method parameters and spectra were validated at 1% FDR using XTandem and Trans-proteomic pipeline (Figure 3.29).

None of the peptides were successfully re-acquired using typical DDA settings. For effective targeting, most 'smart' features needed to be turned off indicating that discovery and targeted methods are practically exclusive. Turning off mono-isotopic precursor selection (MIPS) was absolutely essential suggesting that the real time peak detection algorithm is unable to correctly distinguish most low abundance features. Incorrect real-time peak detection would clearly result in a poor correspondence to a target list comprised of monoisotopic m/z values. This is in keeping with the observed failure to detect the correct monoisotopes even during the discovery DDA SUMOylation analyses. The low resolution preview scan also had a significant negative effect, as did poor management of the MS2 intensity trigger threshold. Best results were found by data independent acquisition (DIA), where MS2 events are scheduled without dependence on MS1 observations. Data independent acquisition is much like SRM in that it requires careful scheduling of events. Precursor feature characteristics like monoisotopic m/z and charge, which are taken for granted during DDA analysis, will also need to be determined retrospectively. Although DIA adds further bioinformatic burden to the process, these results indicate that careful data

management and instrument control will enable the targeting of low abundance features required for future application of this methodology.

DDA/DIA	DDA	DDA	DDA	DDA	DDA	DDA	DDA	DIA
MIPS	on	on	x	x	x	x	x	N/A
CSS	on	x	on	on	x	x	x	N/A
threshold	low	low	low	low	low	high	high	N/A
dynamic exclusion	on	on	on	on	on	on	x	N/A
preview	x	x	on	x	x	x	x	N/A

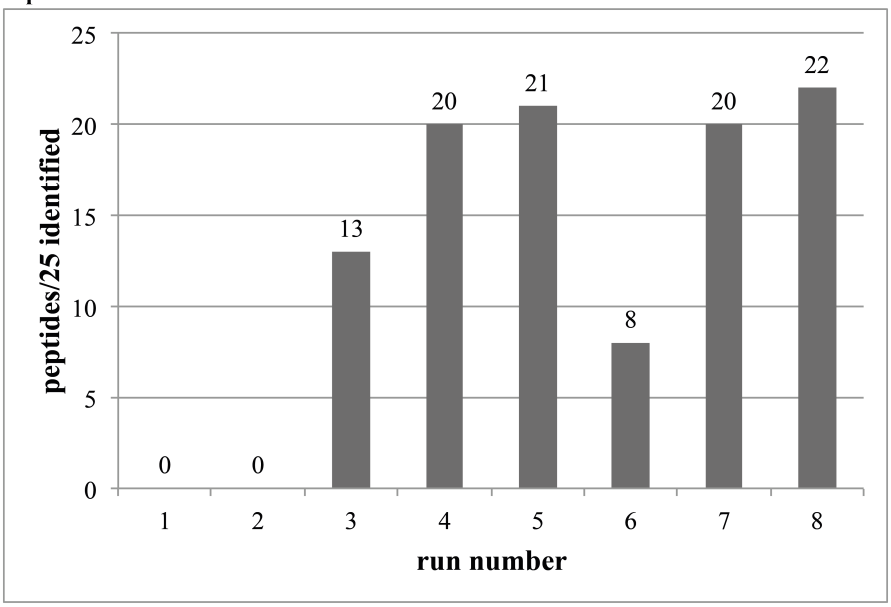


Figure 3.29 MS acquisition parameters have a drastic effect on the success of targeting low abundance peptides.

25 low abundance peptides were selected from a data dependent acquisition (DDA) analysis for targeted analysis. Software features including monoisotopic precursor selection (MIPS), preview scan, sequencing threshold, and dynamic exclusion have drastic effects of the ability to acquire identifiable spectra. Superior results are found by data independent acquisition (DIA).

3.3.7.4 Improved spectral interpretation

Although identifying SUMO isopeptides is a challenging task in *in vitro* scenarios, using isotope coded UBL offers a straightforward approach for confident identification. The well defined simple composition of *in vitro* UBL reactions allowed us to determine isopeptide identities solely based on precursor mass accuracy. This is a luxury not

available in complex samples. Elution time is traditionally an important characteristic for feature identification through accurate mass and time (AMT) tags, but because isopeptides with large remnants all share a common peptide sequences, the substrate peptides cause only a small shift in chromatographic properties. Figure 3.30 shows how isopeptides from the RNMT SUMOylation all elute within a small time range. False assignment to an incorrect peptide is of greater concern for isopeptides with long remnants than for tryptic peptides in general. Interpretation of fragmentation spectra is therefore paramount to confirming UBL substrates in complex samples.

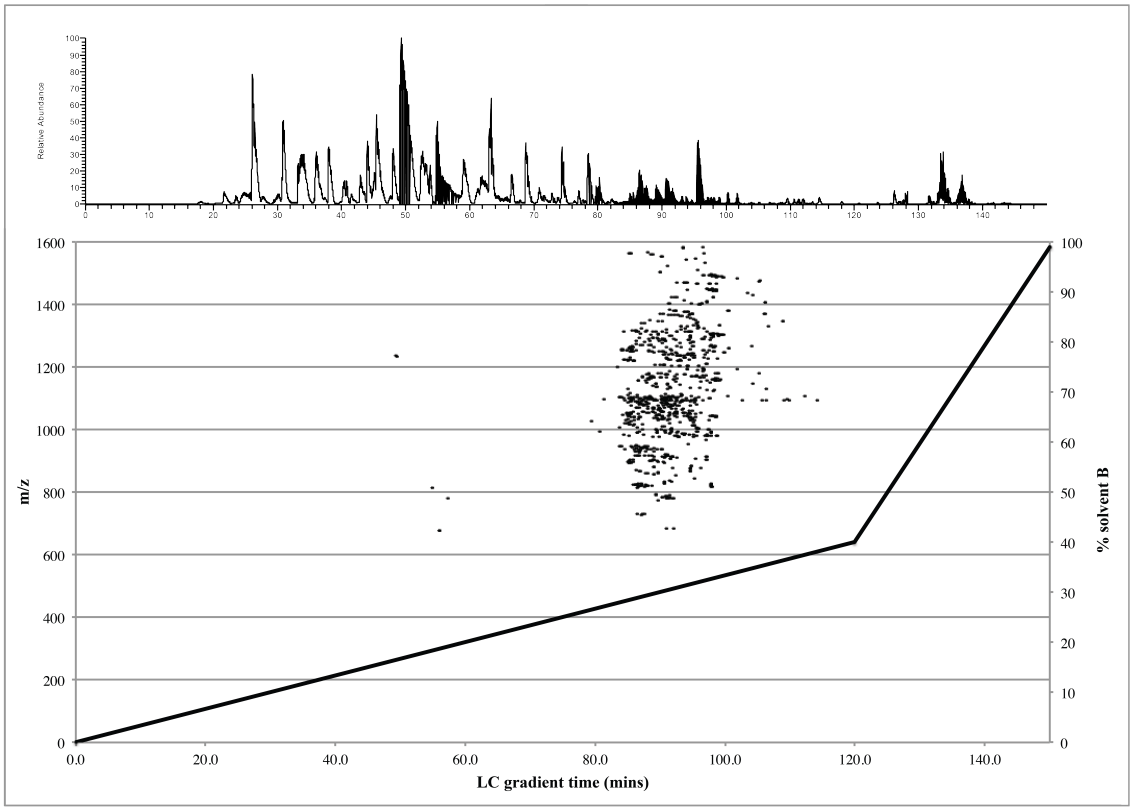


Figure 3.30 SUMOylated tryptic isopeptides have common chromatographic properties.

An ICID analysis of the light/heavy SUMO2 RNMT shows that detected isopeptides (both fully cleaved and miss-cleaved remnant), all elute within a small percentage solvent off C18. Each data point in the scatter plot represents a detected isotopic pair at their elution time and precursor m/z (x and y axes, respectively) and the black line represent the solvent gradient during the LCMS analysis (secondary y axis). A base peak chromatogram above is time aligned with the scatter plot to provide a reference to the whole sample.

The additional use of MS2 spectral interpretation with SUMmOn aided identification confidence for *in vitro* isopeptides through concordance of MS1 and MS2 based methods. Unfortunately, interpretation of low resolution MS2 spectra with SUMmOn is also limited to small databases to prevent false positives (Pedrioli et al., 2006). Greater specificity can be achieved through high resolution MS2 spectra and fragmentation techniques like HCD has also been reported to give better fragmentation than CID for large high charge state peptides (Jeram et al., 2010). Acquiring high resolution MS2 spectra is relatively slow and of low sensitivity in hybrid Orbitrap instruments due to a long flight path through the ion trap. However, more recent Orbitrap designs like the Q Exactive and Fusion offer faster scan rates and significant increase in sensitivity due to the use of quadrupole ion filters.

Interpretation of high resolution isopeptide spectra is unfortunately limited. Techniques like ChopNSpice conveniently utilise capabilities of traditional search engines but result in incomplete spectral coverage over substrate peptides. To discriminate between UBL isopeptides, interpretation of fragmentation over the full substrate is essential. Software support for high resolution MS2 could be implemented by updating the SUMmOn strategy, which already offers theoretically ideal spectral coverage. Alternatively, *in silico* isopeptide fragmentation support could be implemented in an existing open source search engine. An alternative strategy for improved isopeptide spectral interpretation is to utilise the MS3 capability of ion traps. The MS2 fragmentation can create isopeptide y-ions with a shortened remnant while subsequent MS3 can potentially validate the MS2 identification by increasing substrate peptide fragmentation. MS3 has been used to manually validate SUMO isopeptides

(Matic et al., 2008) but here is no software support for automated MS3 isopeptide identification.

Whether moving from *in vitro* to *in vivo* analysis, or directly targeting isopeptides in complex samples, sensitive and specific MSn spectral interpretation for isopeptides is critical for validating substrate peptides. Bioinformatic techniques for isopeptide spectral interpretation is still under developed and investment in research in this area will be essential for successful implementation of *in vivo* UBL isopeptide research.

Chapter IV
HyperProphet

4.1 Introduction

The research presented in the previous chapters was conducted with focus on purification and identification of UBL modifications. The work was necessarily approached from the reductionist perspective to address technical challenges of studying small events that occur within a proteome. In contrast, the work in this chapter approaches proteomics from a systems perspective - the study of the entire proteome. Even in the study of UBL modifications, the quantitative state of the proteome is important for understanding the role of the modification. For example, putative E3 ubiquitin ligase targets can be discovered by finding protein accumulations in E3 mutants (Burande et al., 2009). And while many ubiquitin substrates have been discovered in recent years, quantifying site occupancy, which requires the reference of a total proteome quantification, has yet to be achieved for the majority of ubiquitin targets (Carrano and Bennett, 2013). Global proteomics studies are also necessary to understand the biological significance of perturbations in poorly understood pathways and uncharacterised gene products, such as comparing proteome-wide effects of deleting a panel of DUBs from the yeast genome (Poulsen et al., 2012). The challenge of analysing entire proteomes is a limiting factor in proteome research and the impact extends beyond the analysis of UBLs and PTMs in general. Although the research presented here was initiated with UBL research in mind, because it has a more widespread application in proteomics, its presentation has been abstracted to the study of unfractionated proteomes in general.

4.1.1 Proteome complexity limits proteome analyses

Despite the ever increasing speed and sensitivity of mass spectrometry instrumentation, we are still limited in our proteome analysis capabilities. (Michalski et al., 2011a) points out that the mammalian proteome has 100,000 peptides that are detectable in a typical analysis, yet we achieve identification of only a small fraction of these. In fact, the typical bottom-up workflow where proteomes are enzymatically digested only worsens the complexity. This is unfortunately a necessary step for efficient MS analysis as top-down proteomics on intact proteins has not developed an equivalent level of throughput or sensitivity (Moradian et al., 2014). The requirements for increased peptide coverage for more robust quantification, and coverage in more complex proteomes, still surpasses the ability of current MS hardware to achieve sufficient proteome coverage on a chromatographic time scale. Although hardware and MS acquisition algorithms will no doubt continue to improve, development in this area is out of the hands of most proteomics researchers and we are left to develop alternative means to dealing with proteome complexity.

4.1.2 Overcoming complexity through sample pre-fractionation

With the desire to cover the entire proteome, approaches have been devised to circumvent the MS limitations - the most common of which is sample fractionation. LCMS almost exclusively implies online 1D fractionation with reverse phase chromatography, which is used for its high resolution and compatibility with electrospray ionisation. Some alternative online fractionation techniques have however been investigated for complex mixtures, such as capillary zone electrophoresis (Zhu et al., 2013). Where a single online peptide separation has been insufficient, additional pre-fractionation has been common. Long standing techniques for protein analysis like 1D-PAGE and 2D-PAGE are still commonly used prior to LCMS today (Issaq et al.,

2002). Strong cation exchange (SCX) was also adopted early and has been adapted for automated online 2D fractionation (Washburn et al., 2001). Additional techniques to accompany reverse phase LCMS are continuously being developed, including as high pH reverse phase chromatography (Delmotte et al., 2007) and isoelectric focusing (IEF) (Hörth et al., 2006). New chromatographic media are also still being developed to improve offline fractionation, such as hydrophilic variants of strong anion exchange (Ritorto et al., 2013). More recent examples attempting to expand proteome coverage include extensive fractionation using multiple techniques. Impressive quantitative coverage over the majority of the yeast proteome has been achieved by using solubility fractionation, SDS-PAGE, SCX, and isoelectric focusing (de Godoy et al., 2008; Rezgui et al., 2013) and combinations of fractionations are also used in series for 3D fractionation (Atanassov and Urlaub, 2013). However, sample fractionation has the drawback of creating a significant increase in the samples that require analysing. Given limited MS resources, removing the need for fractionation will enable resources to be applied to increasing replicates for improved statistics and quantitative accuracy, and increasing the biological scope of research projects to include a wider array of biologically informative conditions.

4.1.3 Overcoming complexity without sample pre-fractionation

Analyses on unfractionated samples have the advantage of requiring smaller sample amounts, a more streamline sample processing, and is more amenable to automation. For all these reasons, research is increasingly focusing on maximising protein coverage without pre-fractionation.

4.1.3.1 Improvements in MS hardware

Mass spectrometers are getting faster, more sensitive, and with increasing resolution. A significant component of the advancements in this area have been associated with the release of new and improved MS hardware into the marketplace. The increased scan speed alone has had a direct impact on peptide identification rates in proteomic analyses (Hebert et al., 2014). The release of the Thermo Q Exactive made high resolution mass spectrometry an even more affordable option for many laboratories. Achieving identification of 2500-4000 yeast proteins over 90-240 minute gradients is now achievable on relatively inexpensive MS hardware (Michalski et al., 2011b; Nagaraj et al., 2012; Thakur et al., 2011). On the newer Thermo Orbitrap Fusion, even faster scan speeds enable approximately 4000 yeast proteins in a single 1 hr analysis (Hebert et al., 2014). Extensive fractionation previously revealed that there are at least 4399 of 6400 yeast open reading frames expressed at one time (de Godoy et al., 2008), indicating that we can identify about 90% of the yeast proteome in a short analysis time. This does not imply that we can yet quantify 90% of the yeast proteome. It does however highlight that current and future MS hardware will enable sufficient coverage of proteomes to reveal insights into biological systems without sample fractionations.

4.1.3.2 LCMS acquisition optimisations

The typical reverse phase LCMS with data-dependent acquisition (DDA) has proven to be a valuable means for identifying proteins in both uncharacterised and well characterised proteomes. Although the DDA algorithm is a semi-random peptide selection with little control by the user, minor adaptations can be applied to acquisition parameters to guide DDA towards an improved proteome coverage. The sequence intensity threshold can be optimised such that it is low enough to identify low abundance peptides but above the limits of detection and above an ideal signal to noise

threshold to ensure quality spectra acquired (Wong et al., 2009). Many other parameters are also available to MS operators, including isolation window, injection time, transient time, scan rates, dynamic exclusion, and optimising can have a positive effect on proteome coverage. However, optimising on total protein coverage can potentially be at the expense of sensitivity (Kalli et al., 2013; Kelstrup et al., 2012). Such optimisations are also dependent on the sample and the hardware being used, therefore re-optimisation is required for each independent scenario for optimal performances.

4.1.3.3 HPLC optimisations

Ultra high pressure nano-flow HPLC with small C18 bead sizes has been shown to improve peptide peak capacity, which leads to good peak resolution for long gradients (Cristobal et al., 2012). By improving the chromatographic resolution coupled to MS, the sample complexity at any given time during the gradient is reduced. HPLC gradients can also be modified non-linearly to even out peptide elution rates to minimise regions of high complexity (Pirmoradian et al., 2013). HPLC buffer additives have also been found to improve peptide sensitivity. The addition of DMSO (Hahne et al., 2013) and benzyl alcohol (Li and Li, 2014) were found to improve peptide ionisation and also collapse peptides towards a single charge state (Meyer and A Komives, 2012).

4.1.3.4 Bioinformatic solutions

Bioinformatic solutions are often aimed at overcoming limitations imposed by DDA algorithms, which fail to reproducibly identify peptides within a complex proteome sample due to the semi-random nature of peptide selection. A more targeted approach can be used where a preselected list of peptides is determined for subsequent analyses. This knowledge can be used for directing the MS to target specific peptides from a list

for fragmentation. (Savitski et al., 2010) found that this approach yielded 24% more proteins than DDA acquisitions alone. However, this approach is only applicable if the proteome is already well characterised and a predetermined set of peptides can be selected from preceding analyses. A more generic approach to targeting peptides can be done by creating inclusions lists containing all unidentified chromatographic features observed during initial DDA analyses (Schmidt et al., 2008). Similarly, in a process called post analysis data acquisition (PAnDA), a targeted approach can be achieved by subtracting peptides identified by DDA from the targeted feature lists, and iteratively subtracting identified features after repeated cycles of targeted MS analyses (Hoopmann et al., 2009).

A common theme for each of these approaches is the requirement for repeated LCMS acquisitions prior to or during their bioinformatic applications. From the perspective of maximising MS resource efficiency, these approaches do not necessarily offer great advantage over fractionation as the additional analyses time could otherwise be spent on an equivalent number of sample fractions. Headway was made towards avoiding the repeat acquisition redundancy with the introduction of the "match between runs" feature in the MaxQuant software package. In contrast to previous bioinformatic approaches, this feature retrospectively mines for unidentified features, and instead of repeating a targeted MS analysis, it acquires identifications from other existing replicate or similar analyses for which matching features have been successfully identified. Although this feature has not been formally published, it has been used in a number of publications (Geiger et al., 2012; Nagaraj et al., 2012; Thakur et al., 2011), and highlights the value of acquiring information from similar but non-replicate LCMS analyses.

4.1.4 The future of proteomics

It is perhaps inevitable that biological questions will always outweigh the resources to answer them. But the field of proteomics is not unlike genomics prior to the radical explosion in sequencing throughput that occurred during the human genome project. We are observing rapid developments in MS hardware in terms of capability and value for money, suggesting that the potential for high throughput discovery proteomics may be a reality in the near future. With this in sight, we are conceiving of increasing numbers of biological states to interrogate; screening for disease biomarkers (Pan et al., 2005), proteome-wide screening of protein complexes (Gavin et al., 2006), personalised proteomes (Marko-Varga et al., 2007), and environmental proteomics (Gotelli et al., 2012). To achieve the throughput required, a focus on rapid analysis of unfractionated proteomes is essential.

Most publications announcing unprecedented proteome coverage are boasting protein identification rates. However, identification of a protein is rarely useful in complex proteome analyses, other than to catalogue expressed genes. The overlap in proteins expressed between different cell types and under differing biological conditions indicate that the informative unit of proteomic information is quantitation. It is quantitative proteomics that will enable us to compare proteomic snapshots to elucidate responses to genetic perturbation, drug treatments, disease states, and better understand normal proteome dynamics within healthy cells. Future developments in throughput proteomics will therefore need to be conducted with quantitation in mind and throughput cannot be achieved at the expense of quantitative proteome coverage or quantitative precision.

4.1.5 Objectives of research

Presented is a novel sample processing workflow and bioinformatics solution for the quantitative analysis of unfractionated proteomes. This work was carried out to fulfil the primary aim of extending the limitations of typical DDA LCMS analyses. There were also a set of additional objectives that set the framework for this research. As quantitation was a primary focus, solutions were made to be compatible with isotope labelling, particularly SILAC as it has been observed to provide a higher precision than chemical labelling (Oppermann et al., 2013). Because advances in this domain of research will result from community contributions in multiple academic fields, software was intended to be open source, and kept as operating system and MS vendor neutral as possible. Additionally, software solutions should utilise or interface with existing open source platforms where appropriate. Finally, application of any new developments should ideally not require extensive mass spectrometry or bioinformatic expertise to apply so that solutions can be easily utilised by the general proteomics community.

4.1.6 Biological Application: ELP3

The majority of the work presented in this chapter can be considered from a technical perspective, with the development of new workflows and tools that would benefit a broad range of research that relies on proteomics. To put the work into a biological context, the yeast Elongator complex is used as a subject to exemplify the impact of this research. More specifically, *elp3*Δ yeast will be investigated to better understand the effect *ELP3*-dependent tRNA modifications have on the proteome.

The Elongator complex is named for its association with actively elongating RNA polymerase II (Otero et al., 1999). The Elongator complex is formed through the association of two sub-complexes, Elongator complex proteins Elp1p-Elp3p, and Elp4p-Elp6p (Krogan and Greenblatt, 2001). Its catalytic member, Elp3p, has

acetyltransferase activity and is thought to acetylate nucleosomes, specifically histones H3 and H4, in the path of the elongating polymerase (Winkler et al., 2002). Other potential roles for the Elongator complex are wide spread, including cytoplasmic kinase signalling (Cohen et al., 1998) and exocytosis (Rahl et al., 2005), and deletions of any member of the *ELP*-complex result in salt, caffeine, and temperature sensitivity (Krogan and Greenblatt, 2001). However, the primary role of the *ELP*-complex has since been proposed to be the modification of tRNAs (Huang et al., 2005). Previously proposed roles in protein acetylation are under question as many phenotypes related to previously proposed roles for *ELP* are in fact suppressed by over expression of *ELP* targeted tRNAs in an *elp3*Δ background (Esberg et al., 2006).

The *ELP*-complex is required for formation of 5-carboxymethyluridine (cm5U), which is a precursor to 5-methoxycarbonylmethyluridine (mcm5U) and 5-carbamoylmethyluridine (ncm5U) modifications on uridine-34 of tRNAs. Deletion of any member of the *ELP*-complex abolishes formation of these modifications (Huang et al., 2005). Figure 4.1 provides an overview of the structure of the uridine modifications, and a simplified synthesis pathway (Chen et al., 2011a).

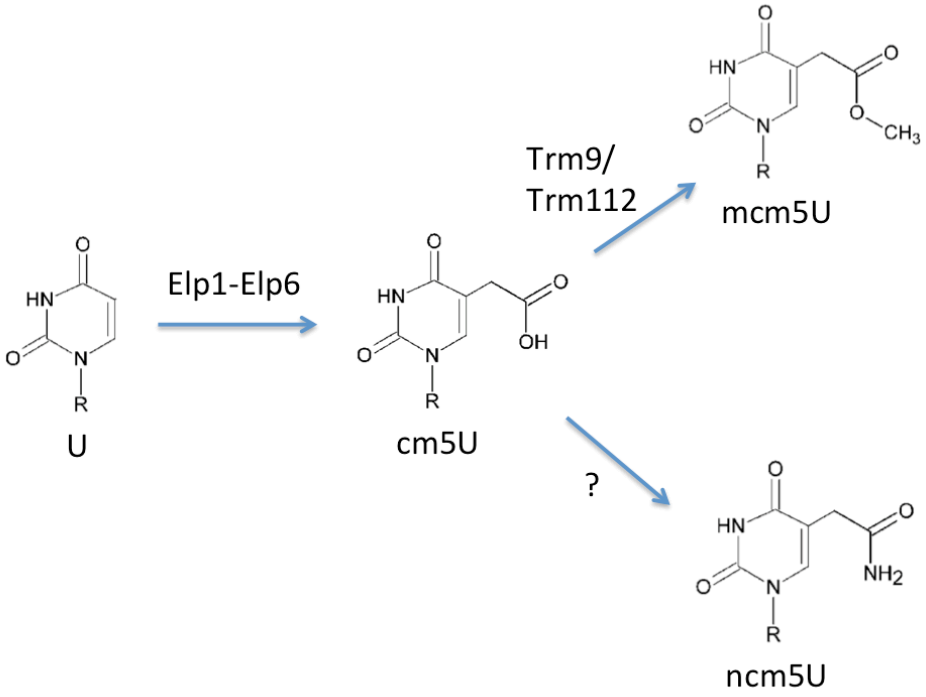


Figure 4.1 A model for the formation of mcm5 and ncm5 uridine.
Structures of uridine, 5-carboxymethyluridine (cm5U), 5-methoxycarbonylmethyluridine (mcm5U), and 5-carbamoyl- methyluridine (ncm5U). The synthesis pathway has not been fully elucidated, but the *ELP*-complex is required for synthesis of the cm5U precursor (Chen et al., 2011a). U = uridine. R = ribose.

The 20 amino acids are coded for by 61 codons (not including stop codons), and most amino acids are coded for by multiple codons. Reflecting this redundancy, tRNAs have the potential to wobble at the third codon position to recognise near-cognate codons as well as their exact matching cognate codon (Crick, 1966). *ELP*-dependent modifications occur on uridine-34 of tRNAs corresponding to 11 codons, which account for almost all of the 13 uridine-34 tRNAs (Table 4.1). These modifications at the tRNA wobble position are important for governing the specificity of codon recognition (Figure 4.2) and can influence the extent of wobble to restrict recognition to the cognate codon, or enhance recognition of near-cognate codons (Agris, 2004). For example, the mcm5U modification on wobble uridines for both tRNA-Arg(UCU) and tRNA-Gly(UCC) are important for reading G in the wobble position (JOHANSSON et al., 2008). Conversely, the ochre suppressor tRNA that reads the ochre stop codon UAA, requires

the mcm5U modification for recognition. Deletion of *ELP3*, and therefore inhibition of ochre suppressor tRNA modification to mcm5U, prevents recognition of the cognate codon (Huang et al., 2005). The mcm5U and ncm5U tRNA modifications are required for accurate and efficient translation and proteins rich in their respective codons are poorly translated in the absence of the modifications (Svejstrup, 2007). In fact, replacing codons for synonymous *ELP*-independent codons can restore protein expression (Bauer et al., 2012).

codon	modification	amino acid	codon	modification	amino acid	codon	modification	amino acid	codon	modification	amino acid
UUU		Phe	UCU		Ser	UAU		Tyr	UGU		Cys
UUC		Phe	UCC		Ser	UAC		Tyr	UGC		Cys
UUA	**ncm5Um**	Leu	**UCA**	**ncm5U**	Ser	UAA		STOP-Ochre	UGA		STOP-Opal
UUG		Leu	UCG		Ser	UAG		STOP-Amber	UGG		Trp
CUU		Leu	CCU		Pro	CAU		His	CGU		Arg
CUC		Leu	CCC		Pro	CAC		His	CGC		Arg
CUA		Leu	**CCA**	**ncm5U**	Pro	**CAA**	**mcm5s2U**	**Gln**	CGA		Arg
CUG		Leu	CCG		Pro	CAG		**Gln**	CGG		Arg
AUU		Ile	ACU		Thr	AAU		Asn	AGU		Ser
AUC		Ile	ACC		Thr	AAC		Asn	AGC		Ser
AUA		Ile	**ACA**	**ncm5U**	Thr	**AAA**	**mcm5s2U**	**Lys**	**AGA**	**mcm5U**	Arg
AUG		Met	ACG		Thr	AAG		**Lys**	AGG		Arg
GUU		Val	GCU		Ala	GAU		Asp	GGU		Gly
GUC		Val	GCC		Ala	GAC		Asp	GGC		Gly
GUA	**ncm5U**	Val	**GCA**	**ncm5U**	Ala	**GAA**	**mcm5s2U**	**Glu**	**GGA**	**mcm5U**	Gly
GUG		Val	GCG		Ala	GAG		**Glu**	GGG		Gly

Table 4.1 Table of codons corresponding to *ELP*-dependent tRNAs.

Codons are tabulated with amino acids and those corresponding to *ELP*-dependent tRNA modifications are in bold. tRNA modifications are listed as 5-carbamoylmethyluridine (ncm5U), 5-methoxycarbonylmethyluridine (mcm5U), 5-carbamoylmethyl-2'-O-methyluridine (ncm5Um), and the double modification 5-methoxycarbonylmethyl-2-thiouridine (mcm5s2U).

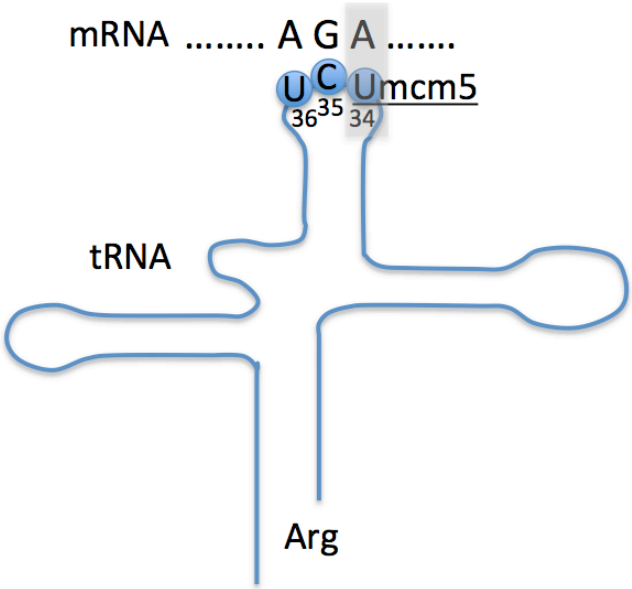


Figure 4.2 Diagrammatic representation of the tRNA anti-codon binding its cognate codon.

In this example, the tRNA-Arg(UCU) recognises its cognate codon AGA. The grey box indicates the site of *ELP*-dependent uridine modification, mcm5U, and is the site of wobble.

In higher eukaryotes, neurological disorders occur if the Elongator complex dysfunction. In humans, mutations affecting expression of IKAP (human ELP1) cause familial dysautonomia (Anderson et al., 2001). It has however been proposed that neural disorders could be caused by defects in the acetylation of α-tubulin by the Elongator complex (Creppe et al., 2009). In Caenorhabditis elegans, inactivation of elpc-1 or elpc-3 cause neurological and developmental dysfunctions (Chen et al., 2009). The role for *ELP* in tRNA modification has only recently been confirmed in mammals, as a mouse Ikbkap/Elp1 mutant exhibits wobble uridine tRNA modification defects (Lin et al., 2013). However, it is still not confirmed whether the Elongator complex involvement in these diseases is due to its role in tRNA modification or direct acetylation of the protein substrates (Chen et al., 2009; Creppe and Buschbeck, 2011).

A subset of *ELP*-modified tRNAs also receive a second modification. The tRNA-Leu(UAA) recognising the UUA codon also has an additional methyl group on the ribose. The tRNAs Gln(UUG), Lys(UUU), and Glu(UUC), are thiolated in place of an oxygen at position 2 on the uridine-34 resulting in 5-methoxycarbonylmethyl-2-thiouridine (mcm5s2U). This modification is catalysed by the ubiquitin-related modifier1 (URM1) pathway. Although *URM1* is a UBL, it has a role as a sulphur carrier in eukaryotic tRNA modification. The effects on proteome composition of deleting *URM1* have been investigated by (Rezgui et al., 2013) who conducted an extensive SILAC proteome analysis of *urm1*Δ vs. wt yeast. Using sample fractionation and 6 biological replicates, in conjunction with an *in silico* codon bias analysis, a correlation between codon biases of AAA, CAA, and GAA, and differential composition of the yeast proteome was observed. This indicated that the lack of tRNA thiolation resulted in impaired translation of a subset of genes rich in the three codons corresponding to the three thiolated tRNAs. Although *ELP3* was also investigated within this paper, an equivalent proteome and codon bias analysis of *elp3*Δ/wt was not conducted.

Although similarities in *urm1*Δ and *elp3*Δ phenotypes have been reported, the *ELP* pathway mediates a far more extensive range of modifications on tRNAs than the *URM1* pathway. In a similar approach to the one conducted by (Rezgui et al., 2013), this chapter presents a differential proteome analysis of *elp3Δ* vs. wt yeast. A codon bias analysis is also used to derive insights into the relationship between *ELP*-dependent tRNA modifications and modulation of the proteome. However, in contrast to the extensive sample fractionation carried out by Rezgui and colleagues, the proteome analysis will be conducted without fractionation and utilising the novel methodologies developed for increasing proteome coverage.

4.2 Materials and Methods

4.2.1 General reagents

Dithiothreitol, Chloroacetamide, $^{13}C_6{}^{15}N_4$-arginine and $^{13}C_6{}^{15}N_2$-lysine, Bicinchoninic acid, formic acid, mass spectrometry grade water and acetonitrile (Sigma). Trifluoroacetic acid, Dimethyl sulfoxide (Thermo Scientific). Sequencing grade-modified Trypsin (Promega). C18 microspin columns (The Nest Group). PicoTip Emitter (New Objective). Magic C18AQ 3 μm 200 Å beads (Michrom Bioresources). RapiGest was synthesised in-house.

4.2.2 Yeast growth media

SD-K-R: 6.7 mg/mL yeast nitrogen base without amino acids and $(NH_4)_2SO_4$ (Becton Dickinson), 0.68 mg/mL Triple Dropout CSM Mix -K, -R, -A (Formedium), 2% (wt/vol) glucose, 0.01% (wt/vol) Adenine, 0.2 mg/mL Pro. Media is supplemented with 0.02 mg/mL arginine or $^{13}C_6{}^{15}N_4$-arginine (Sigma), and 0.03 mg/mL lysine or $^{13}C_6{}^{15}N_2$-lysine (Sigma) for light or heavy SILAC media, respectively.

4.2.3 Cell culture

Wild-type (BY4741 lys1Δ::KAN, lys2Δ::KAN, arg4Δ::KAN) and elp3Δ (BY4741 lys1Δ::KAN, lys2Δ::KAN, arg4Δ::KAN elp3Δ::His3MX) were inoculated into SD-K-R supplemented with arginine and lysine, or $^{13}C_6{}^{15}N_4$-arginine and $^{13}C_6{}^{15}N_2$-lysine. Pre-cultures were grown overnight from a single colony then diluted to OD600 ~0.05 and grown to OD600 ~0.8. Light and heavy cultures were either kept separate or mixed such that OD x volume were equal, then cells were pelleted and frozen in N_2 prior to further processing. For *ELP3* analyses, 3 SILAC mixed cultures were wt-light/*ELP3*-heavy, and an additional 3 label switched to *ELP3*-light/wt-heavy.

4.2.4 Protein and peptide preparation

Cells were lysed with 2 M NaOH, 1 M β-mercaptoethanol for 5 min on ice, then proteins were precipitated with 50% trichloroacetic acid on ice for 10 mins. Proteins were precipitated by centrifugation at 3,000 x *g* for 5 min and washed twice with cold acetone. Pellets were dried and resuspended in 8 M urea, 0.5% RapiGest, 100 mM Tris pH 7.5. Protein content was determined by diluting samples 20x in water, 25 µl of which was mixed with 1 ml bicinchoninic acid, 0.08% $CuSO_4$ and incubated for 1 hr at 37 °C before measuring optical density at OD562. Concentrations were interpolated from a standard curve of BSA. Approximately 40 µg protein was reduced in 5 mM DTT for 1 hr at 30 °C, then alkylated with 15 mM chloroacetamide at RT for 30 min in the dark. Samples were diluted 10x in 50mM tris pH 7.5 and digested overnight at 37 °C with 1:50 sequencing grade-modified Trypsin (Promega). RapiGest was precipitated with TFA, and peptide were purified in C18 microspin columns (The Nest Group). Peptides were dried and resuspended in 0.1%TFA.

4.2.5 Mass spectrometry Data acquisition

1 µg of peptide from *ELP3* SILAC or corresponding single channels were injected onto a ~45 cm self packed 75 µM inner diameter PicoTip Emitter (New Objective), packed with Magic C18AQ 3 µm 200 Å beads (Michrom Bioresources) which was heated to 45 °C. A Dionex Ulitimate 3000 HPLC delivered a 250 nl/min gradient using buffers A (0.1% formic acid, 3% DMSO, 2% acetonitrile) and B (0.1% formic acid, 3% DMSO, 90% acetonitrile) from 1-40% B over 180 min followed by a 40 min wash and equilibration. Data were acquired on a Velos Orbitrap Pro mass spectrometer (Thermo Fisher Scientific) with lockmass off, at 60k Orbitrap resolution, and with a preview scan triggering data dependent acquisition (DDA) of the top 15 precursors above a 500

precursor intensity threshold. Peptides were isolated within a 2 m/z window for fragmentation by rapid scan ion trap CID. 1:1 and 2:1 yeast control data were acquired essentially as above but using a Proxeon HPLC at 200 nL and buffers without DMSO. MS acquisition used a 445.120025 m/z lockmass and the top 12 precursors were fragmented.

4.2.6 Data analysis

LCMS raw data files were converted to mzXML format using ReAdW.exe. MS2 spectra were assigned to peptide identifications by searching against a *S. cerevisiae* protein database (version 2011-02-03 from SGD http://www.yeastgenome.org/) using X!!Tandem release 10-12-01-1 (MPI parallelised version of X!Tandem, http://wiki.thegpm.org/wiki/X!!Tandem). Search parameters included variable +15.994915 Da for methionine oxidation, and multi-static modifications of +57.02146 Da for carbamidomethylcysteine alone or with +8.014199 Da/+10.0082528 Da for heavy lysine/arginine. Peptide precursor tolerance was set to 25 ppm and 0.4 Da for MS2 spectra, and semi-tryptic peptides were permitted up to 2 missed cleavages. Search results were validated using the Trans-Proteomic Pipeline (TPP) v4.6 running on Linux. Peptide assignments were validated using PeptideProphet and InterProphet, and proteins were validated using ProteinProphet and accepted above 1% FDR according to MAYU (Reiter et al., 2009). XPRESS was used for label-free and SILAC peptide quantitation at 10 ppm around the observed precursor mass. Data was further analysed using an in-house MySQL database and R for Linux version 3.0.2. Statistical analysis was conducted by Bayes moderated t-test in the LIMMA package from the Bioconductor project. Codon bias analysis was conducted as previously described (Rezgui et al., 2013). In brief, a random forest analysis was conducted with 1000 trees and seeds randomly selected between 10-99 using the party package (Strobl et al.,

2008). Classification of significantly changing proteins at 1% FDR as up and down regulated was based on the absolute count of gene codons. Randomised analysis was repeated after changing $\log_2$ protein ratios to 1 or -1. Gene Ontology analysis was conducted using FunSpec (Robinson et al., 2002) using a P-value cutoff of 0.01 and using Bonferroni correction. SILAC label switch peptide filtering was conducted using in-house developed R code (described in Result and Discussion section) with a linear $\log_2$ ratio tolerance of 0.5. HyperProphet software was executed with PeptideProphet p-value thresholds of 0.99 (approximately 0% FDR) for chromatographic alignment, 0.85 (approximately 1% FDR) for peptide transfers, feature intensity minimum of 2000 counts, and LCMS analysis time ranges were restricted to 10-190 min.

4.2.7 Software development

Software development was conducted in R for Linux version 3.0.2 and Eclipse 4.3 (Kepler) using Java 1.6 on Mac OSX and run on both Mac OSX and CentOS linux platforms. External JAR libraries used include jopt-simple-4.3.jar (http://pholser.github.io/jopt-simple/), commons-math3-3.0.jar (http://commons.apache.org/proper/commons-math/), and jrap_StAX_v5.2.jar (http://sourceforge.net/projects/sashimi/).

4.3 Results and Discussion

4.3.1 Characterising proteome complexity

To serve as a model proteome of moderate complexity, yeast total proteome extracts were prepared as single channel samples (unlabelled light and SILAC heavy) and as a SILAC mixture at approximately 1:1. While some proteomes may be of greater or lesser complexity, such as mammalian or bacterial proteomes, a yeast proteome was selected due to its relevance to the research presented in this chapter. To characterise the effect the sample complexity has on proteome analyses, a single SILAC 1:1 sample was repeatedly analysed by LCMS. The resulting data was searched against the SGD proteome database using the XTandem search engine, then protein identifications were accumulated into one through nine merged analyses using the trans-proteomic pipeline (TPP). LCMS data was also acquired for fully SILAC labelled heavy and non-labelled light single channel samples.

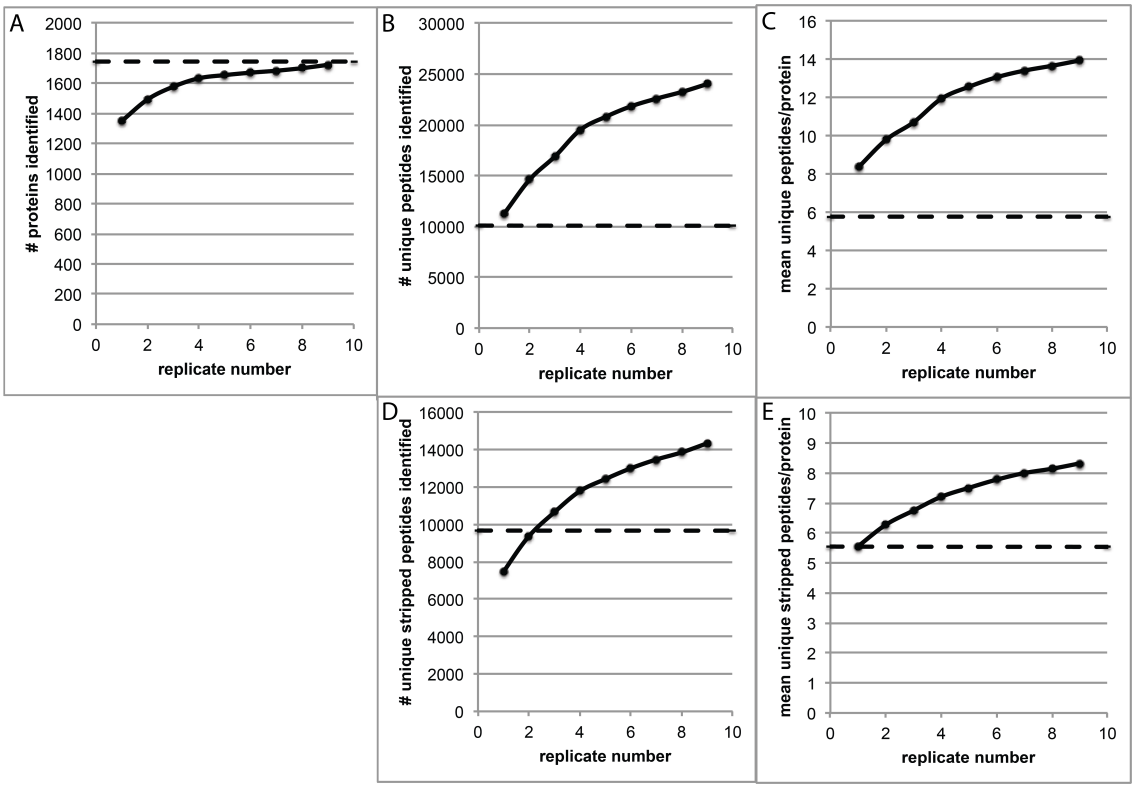


Figure 4.3. Peptide and protein identifications accumulate with repeated analysis.

Protein and peptide identifications were accumulated into one through nine merged SILAC LCMS analyses. A cumulative curve is presented for A) total protein identifications, B) unique peptides (non-redundant peptide sequences with modifications permitted), C) unique peptides per protein, D) unique stripped peptides (modifications ignored), and E) unique stripped peptides per protein. Also included are the mean values for two single channel analyses (dashed lines). All values were extracted from TPP's ProteinProphet and PeptideProphet web interface after filtering at 1% FDR. Note that InterProphet was used to minimise the accumulation of false positives.

Figure 4.3 demonstrates that repeated analysis of a complex mixture yields increasing peptide and protein identifications. The cumulative advantage decreases with increased replicates as the detectable proteome is consumed, although when the accumulation plateau finally occurs is not evident in this data. A further comparison between SILAC and single channel analyses is revealing about the complexity of these samples. More proteins are identified in single channel analyses than in SILAC samples and this holds true even after accumulating proteins over multiple SILAC runs. In contrast to the protein level identifications, SILAC analyses appear to have an advantage for unique peptides identified and peptides per protein. However, considering that peptide modifications are included in the unique peptide count, these numbers are inflated due to the duplication of peptides by the inclusion of heavy arginine and lysine. On considering the unique peptides after stripping of modifications, SILAC and single channel runs have near identical peptides per protein. While the accumulation of unique stripped peptides do overtake numbers observed in single channel analyses, surprisingly, this accumulation contributes more to the number of peptides per protein than the total proteins identified. This observation is consistent with previous yeast SILAC analyses where peptide accumulation also had only modest impact on total protein identifications (Nagaraj et al., 2012).

The discrepancy between single channel and SILAC analyses is noteworthy since they are identical samples other than the SILAC multiplexing. In SILAC labelling, all except C-terminal peptides will incorporate a heavy arginine or lysine, thus the total proteome complexity is almost doubled. Reduce proteome coverage is a consequence of increased complexity. Although SILAC labelling may also introduce a small reduction in sensitivity, as peak heights will also halve at equal protein load, this sensitivity challenge does not appear to become limiting until the plateau begins to occur after the analysis of multiple replicates. The cause of the complexity limitation appears to be largely due to limited MS cycle time, where the number of peptides present simply exceed the sampling rate. However, other contributing factors are related to the nature of DDA, which is a semi-stochastic peptide sampling process. The sensitivity limit, manifested by the plateaux observed in Figure 4.3, may also be the ability of the peptide selection algorithm to penetrate the dynamic range of peptides rather than the true limit of detection of the mass spectrometer. The observed limit of detection can therefore be related to the complexity because peptides of lower abundance may not be selected by DDA in regions of high complexity.

It is important to distinguish detectable peptides, those with features detected in MS1, and identifiable peptides, those that give assignable MS2 spectra. Given that the analysis presented in figure 4.3 is measured by identified peptides and not detected MS1 features, the complexity limitation can also be contributed to by a failure to assign an identity to an MS2 spectrum. This can occur due to co-isolation of two or more peptides within the isolation window of 2 m/z giving a mixed spectrum. This phenomenon may be variable between analyses as the timing of an MS2 event can affect the purity of a spectrum for peptides that are only partially overlapping. Furthermore, increased complexity by SILAC labelling will increase the frequency of

peptide co-elution. While mixed spectra can be assigned to multiple peptides (Wang et al., 2010), this is not achieved by standard proteome software. It should be noted that co-eluting peptides within the 2 m/z MS2 isolation window may have clearly resolvable MS1 precursor ions making these peptides quantifiable at high resolution.

We can conclude for complex samples in general, that any single analysis only identifies a subset of peptides and proteins that are identifiable. Additionally, replicate analyses can improve the total identifications over an experiment. Furthermore, when approaching complexity from a quantitative perspective, it is clear that SILAC multiplexing creates a more complex proteomic landscape that negatively impacts the proteome coverage. Conversely, the lower complexity of a single channel analysis permits a superior dataset for the identification of peptides and proteins. Thus, if we want to benefit from the precise quantitative power of SILAC then we are faced with opposing forces between optimal peptide identification and optimal peptide quantification.

4.3.2 Managing Proteome Complexity: A New Workflow for SILAC Experiments

Given the identification versus quantification dilemma for SILAC experiments, a new workflow is presented to take advantage of both SILAC and non-SILAC samples for the efficient analysis of unfractionated proteomes (Figure 4.4). This proposed workflow is distinguished from a typical experiment in two ways, firstly by the inclusion of single channel samples (either unlabelled light samples or fully labelled heavy samples) providing increased proteome coverage, and secondly, the *in silico* experiment-centric management of peptide and protein identifications.

Obtaining single channel samples requires negligible additional effort as these samples must already be made in the preparation of SILAC samples. Prior to the mixing of light and heavy samples, a portion of each sample should be kept unmixed and further processed in parallel to the mixed SILAC samples. Depending on the experimental design, collection of single and mixed channel samples may occur during cell harvest, or with fully processed light and heavy tryptic peptides. Although the latter requires the least additional processing of samples, mixing SILAC samples as early as possible, i.e. immediately after cell harvest, is recommended to ensure consistent processing of paired channels.

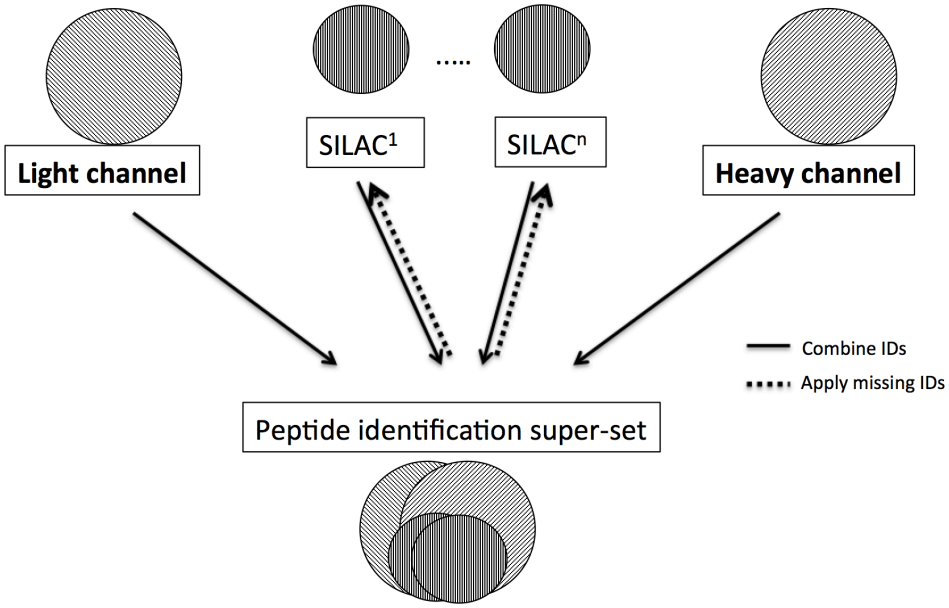


Figure 4.4 A graphical representation of the proposed new workflow for SILAC experiments.

In addition to SILAC replicates, one or more single channel analyses are included in the experiment. All peptides identifications can be accumulated into a single experimental super-set of peptides. Additional peptides are assigned to individual SILAC analyses to aid quantification of peptides that were otherwise not identified.

A wide range of SILAC analyses can potentially be included, such as technical replicates and biological replicates. Similar non-replicate samples may also be included, such as a set of mutants, various environmental stimuli and drug treatments - a prerequisite being that each sample is a quantitate variation on the same proteome. The inclusion of single channel runs is technically optional as this concept is still applicable over just SILAC analyses. However, as outlined in the previous section on proteome complexity, single channel analyses offer increased peptide and protein coverage over a single SILAC analysis. Thus, one or more single channels should be included, for example, a single sample corresponding to a common reference run present in all SILAC analyses. The inclusion of every sample as a single channel is of course possible, although the excessive extra acquisition time on non-quantifiable samples may make this approach impractical.

The decoupling of identification and quantification analyses also permits MS acquisition to be optimised for their respective purpose. Settings such as mass resolution, preview scan option, MS1 duty cycle length, MS2 trigger intensity threshold, and DDA versus targeted acquisition, can dramatically affect search results (Kalli et al., 2013). Furthermore, optimal setting can be contradictory for identification and quantification. For example, a long cycle with many MS2 events may give optimal proteome coverage but insufficient points across a chromatographic peak for accurate quantification. SILAC samples can therefore be acquired with optimal quantification at the sacrifice of MS2 peptide identification scans knowing that adjacent single channel analyses can provide the necessary proteome coverage.

4.3.3 HyperProphet: a software module enabling mixed single-channel/SILAC experiments

The major component of the proposed workflow entails a bioinformatic solution to the redistribution and quantification of peptide identifications. The software solution presented here is called 'HyperProphet'. The design of this software module set out to fulfil a number of aims; to be operating system independent, MS vendor independent, to be open source, to be modular, and to permit interfacing with multiple proteomics pipelines. The Trans-Proteomic Pipeline (TPP), a popular open source proteomics data analysis platform from the Institute for Systems Biology (Keller et al., 2005), was selected as the initial interfacing pipeline as it shares similar philosophies; it is open source, has linux and Windows distributions, is MS vendor neutral, and is built from modular tools. TPP also supports multiple search engines and includes all the necessary tools for peptide validation, protein inference, and peptide and protein SILAC quantitation. Specific examples discussed in this chapter and in appendices will refer to TPP on a linux platform.

HyperProphet is a command-line interface application developed in Java (an operating system independent language). To keep the software modular (having no dependence on other software), the interface to HyperProphet was designed to be solely through the mzXML (Pedrioli et al., 2004) and pepXML (Keller et al., 2005) file formats (Figure 4.5). This enables any proteomics pipeline to potentially interface with HyperProphet, as long as there is compatibility with the xml formats. HyperProphet operates independently from MS vendors, or the search engines used, which is a flexibility that is acquired from the xml formats being used. The mzXML is a standardised, open format for representing MS data. Vendor supplied and third-party converters are available to generate mzXML from vendor specific binary formats

(http://www.proteomecenter.org/software.php). The pepXML format contains peptide level information, such as peptide-spectrum-matches (PSMs), and supports multiple search engines including Mascot (Perkins et al., 1999), XTandem (Craig and Beavis, 2004), SEQUEST (Eng et al., 1994), and Comet (Eng et al., 2013). The pepXML format can also contain additional information including statistical validation of assignments and quantitative information (Keller et al., 2005).

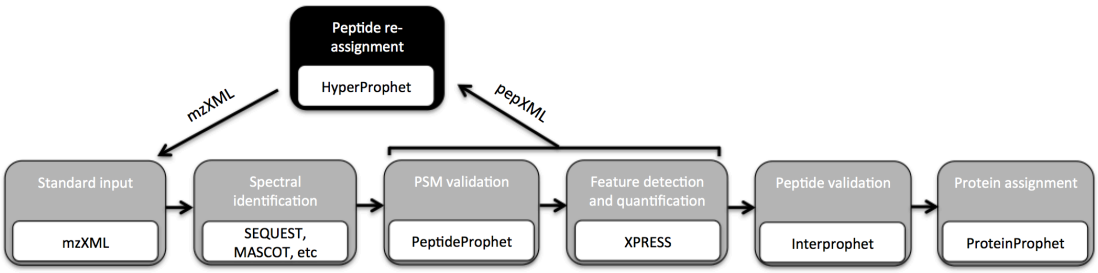


Figure 4.5 HyperProphet interfaces with the Trans-Proteomic Pipeline (TPP)
The HyperProphet module (black) interfaces with TPP modules (grey) though the open xml formats mzXML (representing MS data) and pepXML (representing peptide assignments, validation, and quantification information).

Although the ultimate purpose of HyperProphet is to enable improved quantification of SILAC analyses, HyperProphet does not perform quantification. The specific function of hyperProphet is to manage the collection of peptide identifications over an experiment, and to re-assign peptide identifications when missing from individual runs. The transfer of a unique peptide identification is used to direct the existing tools to perform the quantification. Note that in the context of this work, a peptide refers to a unique LCMS feature identification, thus multiple charge and modification states are considered as unique peptides. The peptide transfer is conducted by generating a new instance of an mzXML file. This new mzXML file is a composite file containing a duplicated set of MS1 spectra and a set of MS2 spectra derived from adjacent runs. This file therefore contains identical quantitative information, but unique identifications. The transfer of MS2 spectra is carefully managed with chromatographic time alignment

and extracted ion chromatograms to ensure that MS2 spectra are inserted within the parent feature elution profile. Each input SILAC analysis in the experiment will therefore receive an additional composite mzXML file. Note that the original data file is left unmodified, making the native and the transferred peptide identifications easily distinguishable by file origin.

Note that future references to HyperProphet analyses necessarily implies the use of the TPP pipeline as a foundation, and a TPP analysis to be an equivalent analysis without the additional use of HyperProphet. To understand the full process (Figure 4.6), the inner workings of HyperProphet will be described before discussing the interaction with TPP.

4.3.4 Hyperprophet algorithm outline

The internal functions of HyperProphet can be divided roughly into four main tasks; accumulation of high confidence peptide identifications, time alignment of LCMS analyses, assignment of missing peptides, and creation of composite mzXML files (Figure 4.6, grey box). A high-level overview of the key decisions and steps HyperProphet executes follows.

Accumulation of high confidence peptide identifications:

- read in each pepXML file, filtering on a minimum input peptide probability
- remove duplicate PSMs
  - duplicates assigned to ‘shoulder’ chromatographic features of lower intensity are discarded
  - preference then given to highest probability and being assigned to a chromatographic feature
  - alternative modifications and charge states are preserved as unique chromatographic features
- tabulate unique peptides into an m x n array (m peptides x n analyses)
  - rows represent common peptides between analyses, null values indicate missing peptides

Time alignment of LCMS analysis:

- for each pair of analyses, collect high probability peptides in common
- create time alignment between runs using a segmented linear alignment
    - This alignment model will be used to determine insertion points for unique peptide transfers

Assignment of missing peptides:

- for each SILAC analysis, determine set of missing peptides
- for each missing peptide, select best spectrum from adjacent analyses
    - preference then given to highest probability and being assigned to a chromatographic feature
- update MS2 retention time to that of the recipient analysis using chromatographic time alignment model

Creation of composite mzXML files:

- For each MS2 to be inserted, create an XIC around expected retention time in recipient mzXML file - refine insertion time to correspond to precursor peak apex
- for each SILAC analysis, create a new mzXML file with duplicated MS1 spectra and insert MS2 spectra at updated retention time

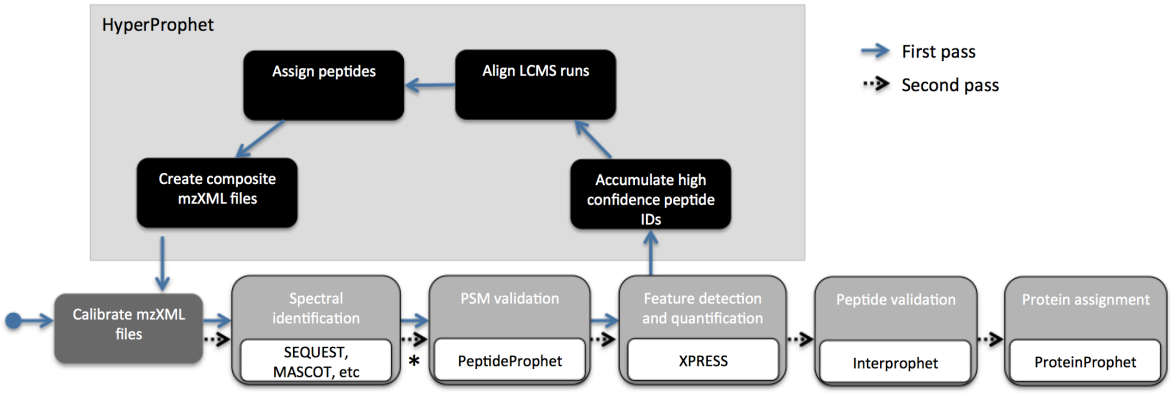


Figure 4.6 HyperProphet and interaction with TPP

The workflow begins (solid circle at left, following solid arrows) with the optional re-calibration mzXML files, followed by a first pass through TPP up to the peptide level validation and label-free quantification for all files in the analysis. HyperProphet is then executed on the set of pepXML and mzXML files generating new mzXML files. The new mzXML files are then recalibrated and passed through TPP again (dashed arrows), including protein validation and quantification. * denotes the point at which native analyses and their HyperProphet analysis partner files are merged into a single result.

4.3.5 HyperProphet peak selection and peak detection

Accuracy of HyperProphet is partly dependent on the quality of peptides selected for transfer. Although only the MS2 spectra are transferred, identifications assigned to quantitative features are preferred to ensure they are bona fide peptides rather than incorrectly assigned noise. Redundant peptide identifications are frequently encountered, most of which are assigned to the same precursor feature, but some are incorrectly assigned to low abundance features. These incorrect features are detected due to erratic signals forming peak shoulders or additional small features in chromatographic tails. As a proxy for feature detection correctness, features are plotted as peak area over intensity. Figure 4.7A displays abnormal features that have unexpected peak area-intensity relationship which occurs mostly at low intensity. The abnormal features are discarded by HyperProphet during the selection of high quality non-redundant peptide identifications (Figure 4.7B). For comparison, features detected in recipient analyses for peptides transferred by HyperProphet display a similar level of quality (Figure 4.7C).

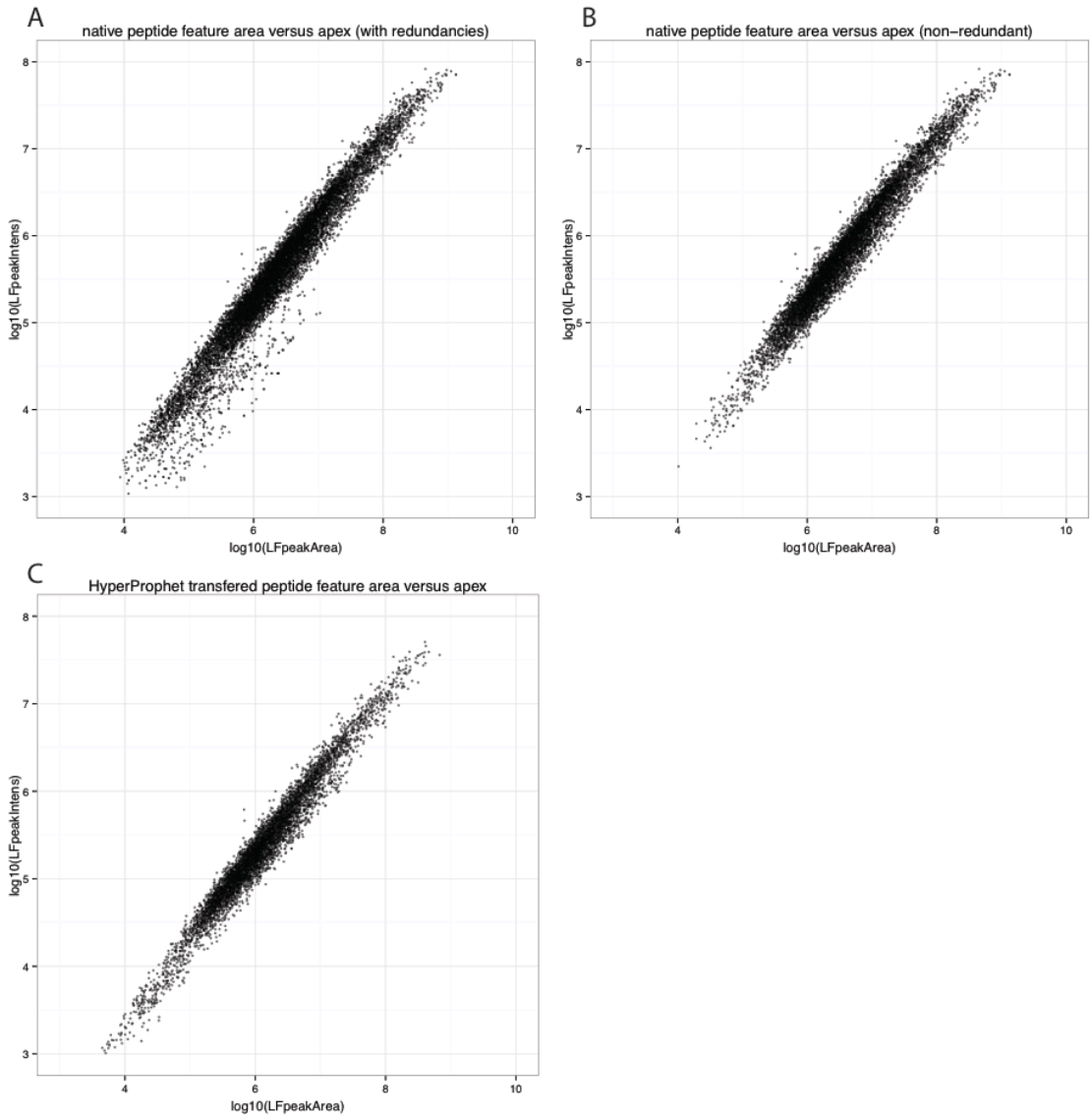


Figure 4.7 HyperProphet transfers peptides from high quality chromatographic features.

The correlation between feature apex and area is well correlated and used as a proxy for peak detection correctness in a yeast SILAC LCMS analysis. The set of all features detected for high probability peptide identifications (A) have outliers at the lower intensity range. These data points are mostly eliminated after selection of high quality non-redundant peptides (B) and upon transfer to a recipient analysis (C).

It is critical that HyperProphet generated mzXML files place MS2 spectra within the MS1 chromatograph peak profile, as this would necessarily be the case for a DDA LCMS analysis and is expected by MS1 quantitation software. To achieve the feature quality indicated in Figure 4.7C, HyperProphet relies heavily on the chromatographic

alignment. An alignment model is constructed using the apex elution times of chromatographic features for peptides in common between each analysis (using MS2 spectra times is also a software option). An example of a replicate analysis shows that there is a very strong correlation (Pearson score R^2=0.9986) between analyses with consistent chromatography (Figure 4.8). Although the alignment could be modelled with a single linear fit in this case, to adjust for variation during analyses, HyperProphet uses a series of contiguous linear fits with outlier removal. Outliers are defined by a user set standard-deviation threshold.

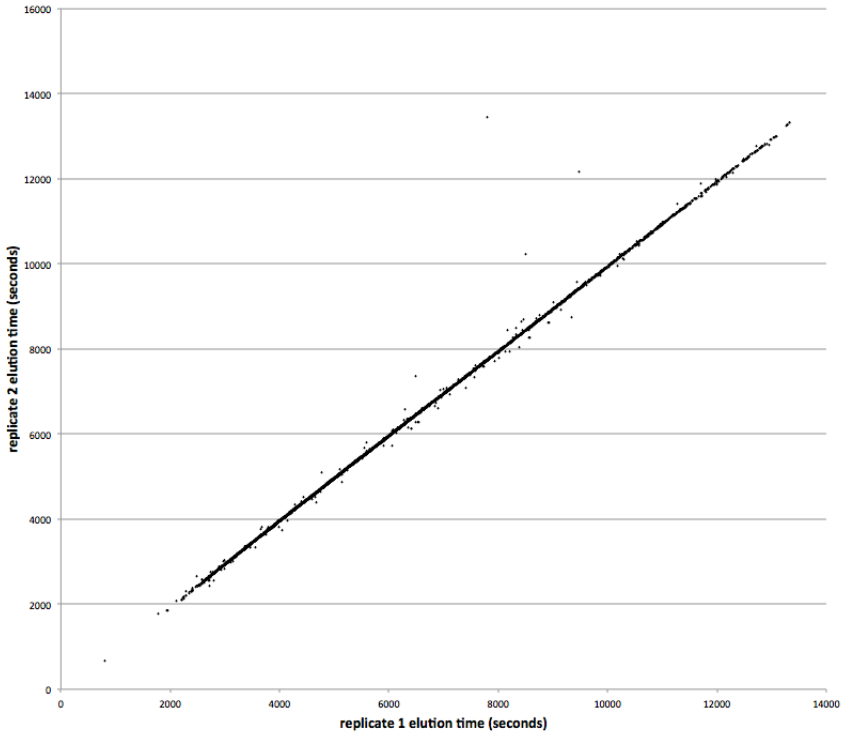


Figure 4.8 Peptide elution times between replicate analyses are highly correlated.

Peptide feature apex elution times are aligned for two technical replicates of SILAC yeast LCMS data filtered at high probability (p>0.99). The alignment model between replicates aids in the prediction of elution times for peptides missing from one replicate.

To further fine tune the insertion time of MS2 spectra, extracted ion chromatograms (XICs) are determined for mono-isotopes of each transferred peptide in the recipient

MS1 scans. The initial insertion time, derived from the segmented linear alignment model, is adjusted to that of the observed peak apex using a greedy hill climbing algorithm. Although this does not guarantee to find the true maximum, a local maximum is sufficient to ensure the insertion point is within the feature profile range. Figure 4.9 demonstrates the effect of the XIC mediated insertion time adjustments. Time adjustments are predominantly around zero seconds indicating that the initial alignment was precise and little or no adjustment is needed. In the 3hr LCMS analysis presented, +/- 30 seconds is ample tolerance for peak adjustment. The adjustment of precursor intensity (figure 4.9B) reveals that the XIC time adjustment does not alter many peptides (y=x correlation) and it improves apex detection for many peptides (scatter of increased intensity). For the remaining peptides, the adjustment is essential due to initially missing the peak (x=0).

As a final confirmation of peak detection accuracy, precursor intensities were compared to feature apex intensities from the donor analysis. A strong correlation is expected because the light and heavy donor analyses are essentially equivalent to the two channels of the 1:1 SILAC recipient, although some variation is expected as each sample represents an independent tryptic digest. Comparing precursor intensity before and after XIC time adjustment (Figure 4.9C-D) shows that spectral insertion vastly improves. A small portion of outlier data points appear to have found a local rather than global maximum intensity.

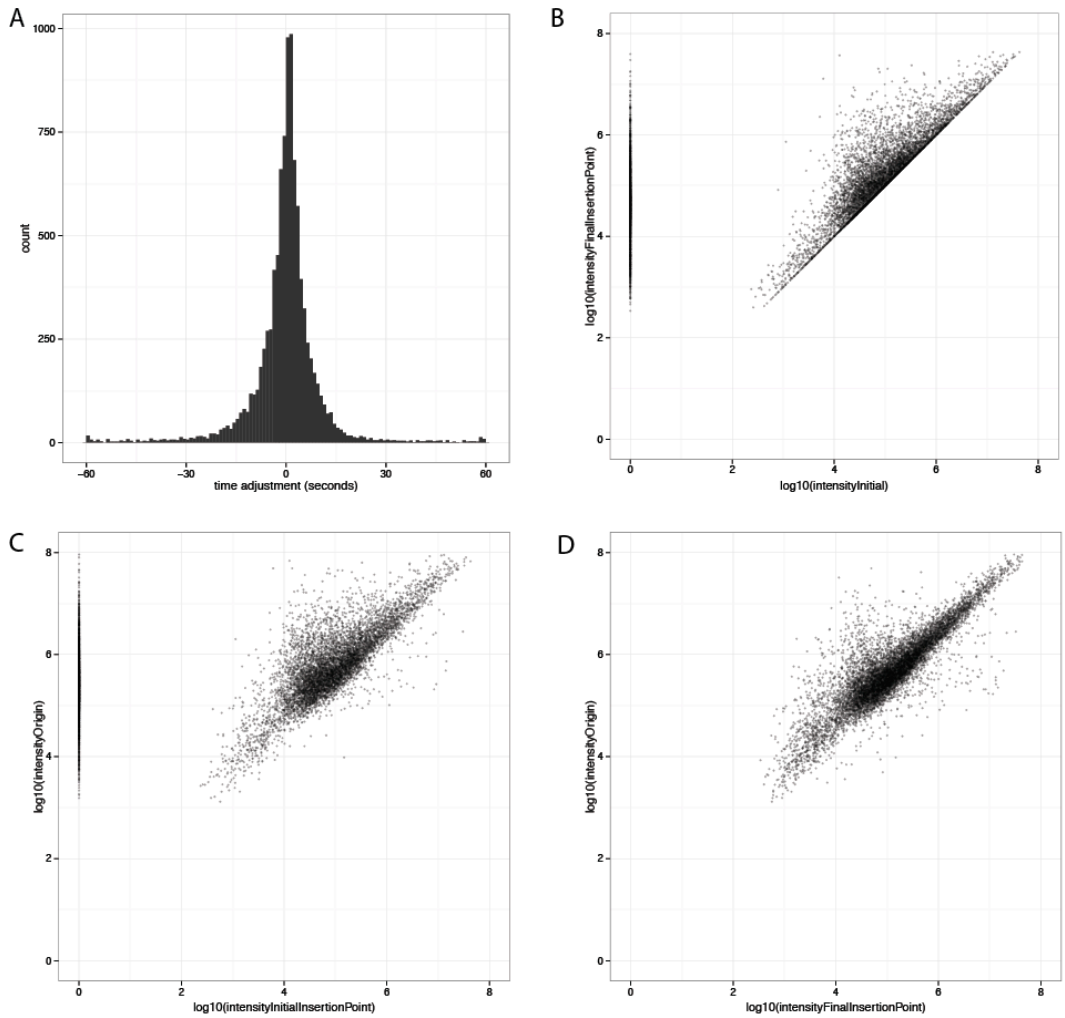


Figure 4.9 Refinement of spectral insertion time improves peak detection and quantification accuracy.

Chromatographic peak detection metrics are presented for a HyperProphet transferring peptides from light and heavy single channel analyses to a 1:1 SILAC analysis. A greedy hill climbing algorithm starts at chromatographic alignment time and then finds a local maximum XIC apex within donor analyses. Only a small time adjustment is required to find an XIC apex (A). Comparing initial and adjusted precursor intensity indicate an improved apex detection for a subset of peptides (B). Precursor intensities are plotted against the observed feature intensity in the donor runs single-channel analyses. The expected correlation is observed between similar donor recipient intensities (C) which improve after XIC adjustment (D). Note that the donor intensities are approximately twice the intensity ($\log_{10}(2)$=~0.3) of SILAC precursor intensities, which is expected due to SILAC signal division.

4.3.6 Mass calibration of LCMS analyses

During the development of HyperProphet, it became apparent that there was a discrepancy between precursor masses of transferred MS2 spectra and the observed

precursor in the preceding MS1 scan in recipient analyses. This inconsistency occurs because mass calibrations can vary between analyses, and the original precursor mass remains associated with the transferred MS2 spectrum (mzXML stores the precursor with the MS2 information). This presents complications for quantification in HyperProphet analyses because peak areas are extracted using MS2 precursor masses from non-corresponding MS1 data. Two potential remedies present themselves; to correct MS2 precursor values to the observed precursor in preceding MS1 spectra, or to fully recalibrate the LCMS analysis such that all mass values fall within a small mass error tolerance.

A simple tool was implemented to correct MS2 precursor values to that of the observed MS1 value. Although this tool was developed with the intent to be applied to HyperProphet generated mzXML files, it was in fact useful for native files which also contained precursor discrepancies. This surprising observation occurred only in analyses acquired with a preview scan (refer to figure 1.11 for further explanation of preview scans). Although preview scan data is not stored by Orbitrap control software, it was reasoned that the MS2 precursor values were derived from the low resolution preview scan and not from the high resolution MS1 scans. Data acquired without a preview scan were not affected by discrepancy between observed MS1 and reported MS2 precursor. The resulting poor mass accuracy affects both database search identification rates and quantitation accuracy. Further investigation into LCMS file calibration error revealed that many analyses had very poor calibrations, even when a lockmass was used (an internal calibrant, typically a polysiloxane at 445.120025). Furthermore, using DMSO as an HPLC buffer modifier suppresses polysiloxane ionisation and does not provide a reliable alternative lock mass. A simple correction for

MS2 precursor masses was found to be insufficient in these cases highlighting the need for visual inspection of data quality and full recalibration of LCMS data.

Some calibration tools already exists part of other pipelines, such as mMass (Strohalm et al., 2010), OpenMS (Kohlbacher et al., 2007), MSinspect (May et al., 2009), and MaxQuant (Cox and Mann, 2008), but none were deemed suitable due to insufficient control over calibration, incompatibility with mzXML file format, operating system platform dependence, or lacked tools for visual inspection of calibration status. As a result, the calibration software was developed into a stand-alone tool for re-calibration of mzXML format LCMS. The tool is developed in Java, uses mzXML and pepXML input formats, and generates a calibrated mzXML output in addition to log files providing data on the calibration results. Viewing calibration results is done graphically calling R scripts. A general overview of important features are discussed here and full options and execution details can be found in Appendix C.

The simplest calibration option is MS2 precursor correction. As already discussed, this option looks for the correct MS2 precursor ions in the preceding MS1 scan and assigns this as the MS2 scan. This option is essential for correcting precursor masses for HyperProphet-transferred MS2 spectra, and for improving precision for spectra triggered by low precision preview scans. Although this option does not recalibrate m/z or time domains, it does increases the mass precision of a dataset and improves peptide identification rates. An option to remove spectra without detectable precursors in the preceding MS1 is also available, which effectively removes non-quantifiable data that occurs within the low signal noise range.

A more advanced feature is the recalibration of m/z and time axes. Deviations over time can be caused by temperature fluctuations and deviations over the m/z axis may also be observed due to detector manufacturing defects. Also note that the use of a lock mass does not correct for m/z distortions. This option imports peptide identifications from pepXML files, with filtering on probability and intensity thresholds. Peptide mass errors are determined for identified peptides (theoretical minus observed mass). Alignments are determined for mass error over time and mass error over the m/z axis using a segmented linear aligner (multiple contiguous lines of best fit) and the data is adjusted to minimise the ppm error based on the alignments. Although this iterative process requires an additional database search after calibration, it offers greater calibration precision and the ability to compare the data quality before and after recalibration. Knowledge of the data precision allows informed decision on the mass tolerances to use for downstream quantitation.

The most powerful calibration comes from single scan calibrations which independently calibrates each individual MS1 scan. This option compensates for erratic fluctuations between scans that cannot be calibrated using a global fit. As this requires a larger dataset to ensure each individual scan has sufficient data points, feature detection is usually required to extract data points across each identified chromatographic peak. This calibration function also creates a segmented linear alignment, but over the m/z domain for each spectra. In cases where there are insufficient data points, calibration reverts to global m/z-time calibrations. This approach has the added advantage of effectively calibrating even when abrupt changes in calibration occur due to errors in lock mass calibration (Figure 4.10).

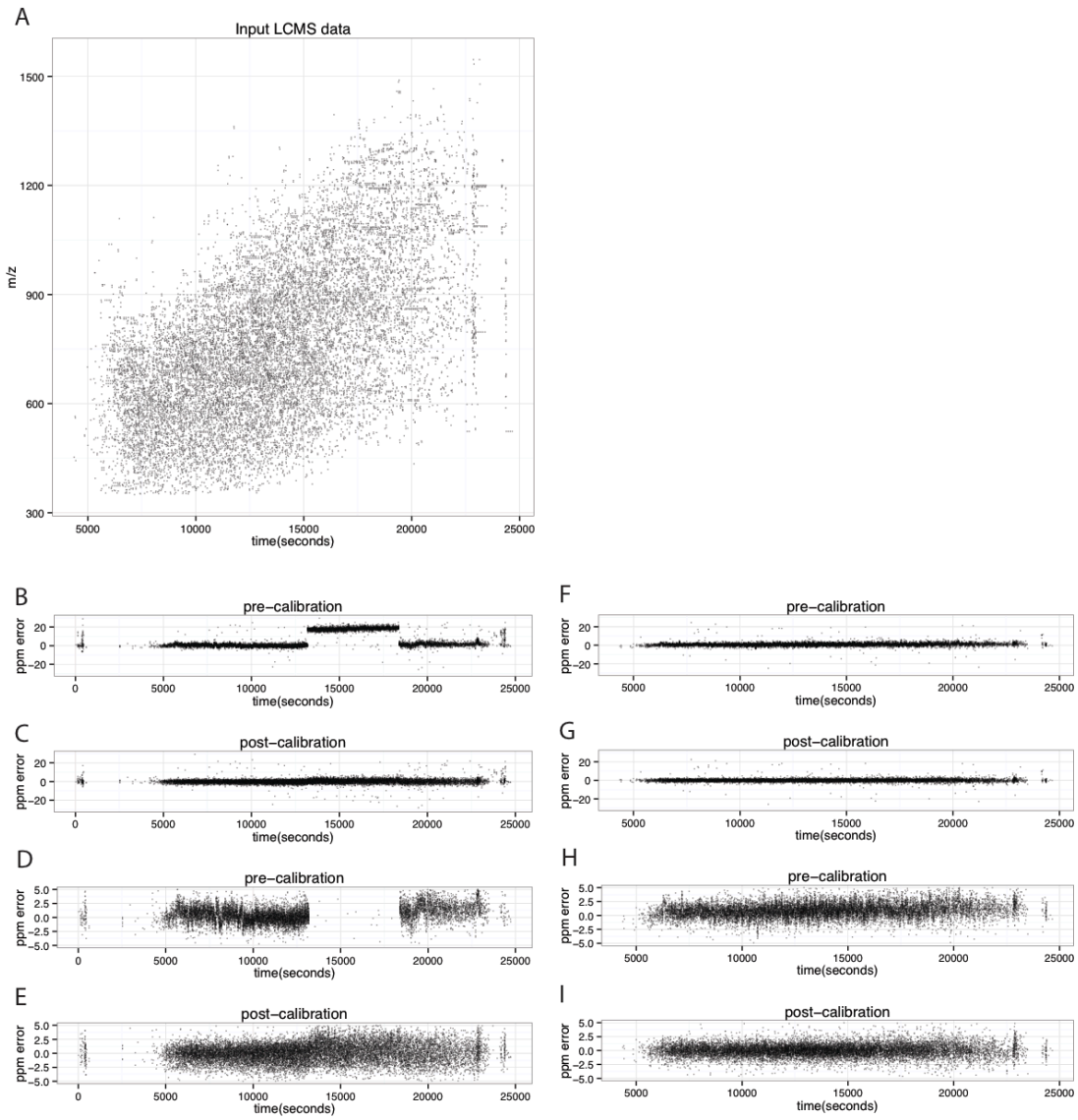


Figure 4.10 Examples of LCMS re-calibration

(A) An example of input peptide data (m/z and elution time for peptides with probabilities >0.99) used for single scan calibration. B-E) An example of an LCMS analysis where the lock mass calibration has failed with catastrophic consequences for quantification. PPM mass error over time can be observed at overview and zoomed ppm ranges (B,D). The mean, standard deviation (stdev), and root mean square (RMS) errors are vastly reduced after calibration and can be visually inspected for improvement (C,E). An improvement re-calibration can also be observed for analyses that are already well calibrated by acquisition-time lock mass calibration (F-I). Note that minimising RMS error, rather than just the standard deviation, is important for quantitative tools that use a symmetrical mass tolerance around the theoretical peptide mass.

4.3.7 HyperProphet: usage and interaction with TPP

Details on the use of TPP can obtained from the website wiki (http://tools.proteomecenter.org/wiki) and publications (Pedrioli, 2010). An overview with detail on steps that are critical for hyperProphet are discussed here, referring to Figure 4.6 for graphical overview. Although the following individual steps appear numerous, they can be grouped into just a few commands by bash scripting and by using TPP's xinteract command wrapper.

The initial steps prior to HyperProphet execution include converting LCMS data to mzXML, optional mass recalibration, protein database search (with output to pepXML format), and label free feature detection using XPRESS. HyperProphet is executed, taking as input an mzXML and pepXML for each analysis located in the current directory. An example execution on a linux/Mac terminal might be:

```
jHyperProphet -d ./ -a 0.99 -o 0.85 -s fileX,fileY
```

where ./ denotes the current directory (containing all input data), peptides are accepted above p=0.85, only high confidence peptides above p=0.99 are used chromatographic alignments, and output is requested for files X and Y (implying that the non-specified files are single channels and do not require output). Full parameters options and descriptions can be found in Appendix D.

Output files are named fileX_HP.mzXML, making this a partner file to the input fileX.mzXML. Additional files are also output with information on the parameters used and a process report (info.txt), a table of all non-redundant peptides imported over the entire experiment (peptideTable.txt), pairwise LC alignments (align-fileX-fileY.txt), and a record of peak detection in each recipient file (targetXICs_fileX.txt). These

additional files can be ignored during the analysis but allow the user to later interrogate the results for performance and accuracy. Further optional re-calibration of mzXML files can occur here to correct MS2 spectra precursors to those found in the MS1 scans - recalling that the transferred spectra did not originate from these MS1 spectra.

The aim of the final pass through TPP (Figure 4.6, dashed arrows) is to merge each pair of mzXML files. The new mzXML files are first re-searched against the protein database to produce pepXML files. Although these individual spectra have been searched before within other files, changes in precursor mass can have a subtle effects on the dataset. Because the HyperProphet files contain a low number of high quality MS2 spectra, this search is very rapid. PeptideProphet is used to merge the pepXML datasets into a single analysis. This merging process is a typical step for fractionated sample data, although in this case the two datasets are the original pepXML and the new HyperProphet derived pepXML. The final steps are as for a typical TPP analysis, including SILAC quantitation with XPress, further statistical peptide validation with InterProphet (which provides better management of false positives), and protein inference and quantitation with ProteinProphet.

4.3.8 Validation on SILAC samples of known protein ratios

It is important to validate the performance of HyperProphet on samples of known composition before applying it to biological samples of unknown state. To confirm that peptide ratios are preserved after transfer by HyperProphet, SILAC light and heavy yeast were cultured under identical conditions and mixed at approximately 1:1 ratio based on culture optical density. A HyperProphet analysis was performed on two replicate LCMS analyses of the 1:1 SILAC sample, with unique peptide identifications

being transferred bi-directionally. To assess the quantitative accuracy of this analysis, resulting peptide and protein quantitations were compared to the quantitative values derived from a TPP analysis without the use of HyperProphet (from the 'first pass' analysis represented in Figure 4.6). Since it is not the task of HyperProphet to validate peptide identifications, analyses were done at a p-value corresponding to approximately 0% FDR to remove the majority of false positive peptide identifications. Proteins were also filtered at approximately 0% FDR and single peptide hits were permitted.

Figure 4.11 compares the SILAC ratios for peptides that are common to both replicates and peptide ratios for unique peptides that are transferred by HyperProphet. The ratio distributions of peptides are tightly clustered at 1:1 (0 in $\log_2$ space) and similar distributions are observed between HyperProphet and TPP quantifications. Although not all peptides are in agreement between replicates, outliers are observed for both TPP and HyperProphet analyses indicating that the majority of these errors occur as a result of replicate variation rather than incorrect peptide transfer. Many peptides are observed towards extreme ratios (infinite ratios are represented at -/+10). These outliers could be due to quantitative errors as a result of interfering signals or false positive identifications with incorrect arginine/lysine composition. While it is outside the purpose of HyperProphet to correct for these errors, it is important to observe that these ratios are preserved between replicates, as indicated by the correlation at x=y.

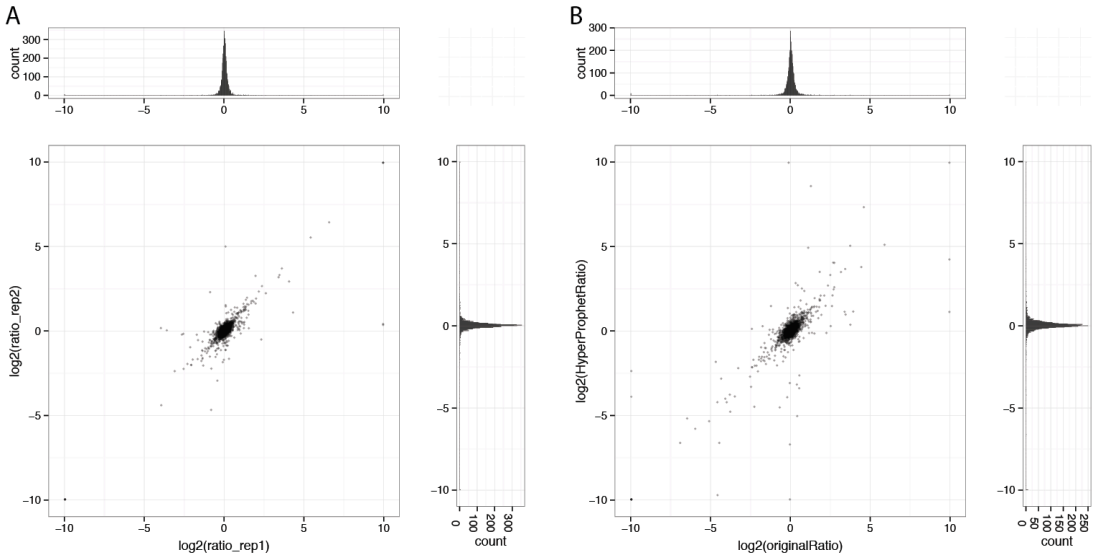


Figure 4.11 HyperProphet preserves SILAC peptide ratios between replicates

In a test case between two 1:1 SILAC tryptic yeast replicate analyses, peptide ratios are well conserved between replicates as indicated by SILAC ratios for peptides in common between replicates (A). When using HyperProphet to transfer unique peptides (B), the ratio of transferred peptides in their recipient analysis (y axis) correlate well to the original SILAC peptide ratio in the donor analysis (x axis). To account for data saturation at 1:1 in the scatter plots, histograms over the x and y axis are represented above and right of each scatter plot, respectively.

To further assess the effect of HyperProphet, a 1:1 SILAC replicate dataset was analysed for protein ratios before and after receiving peptides from the second replicate. Figure 4.12 reveals a similar outcome, where ratios are preserved for the vast majority of proteins. Most of the transferred peptides contributed to existing proteins rather than to new proteins, which is consistent with observations made for protein identifications in figure 4.1. Some new proteins have been added though (Fig 4.12 D) and these proteins will be predominantly composed of HyperProphet transferred peptides since they were not present in the original TPP dataset. The majority of new proteins also clusters at the expected 1:1, although outliers appear more conspicuous due to the low count in this histogram. The outlier at a ratio of +2 (labelled with * in Figure 4.12 D) gives some concern for the potential introduction of errors. On inspection of this outlier, this protein (YPL225W) is composed of only two peptides; one peptide

(K.KIEDYNFGTLLR.T) giving an exact 1:1.00 ratio and the other (K.HGLNDWIVGQK.-) an anomalous 1:0.06. However, the anomalous peptide also had an extreme ratio in the donor analysis (1:0.16) indicating that this peptide was likely to be a false positive and was not incorrectly quantified. The propagating of such quantitative errors is in fact the correct behaviour to expect from HyperProphet because the HyperProphet module does not govern peptide validation or peptide quantitation. It does however highlight a general concern for the quality of SILAC quantitative data, particularly for proteins with very low peptide coverage. Additional control of quantitative outliers will be discussed further in the next section.

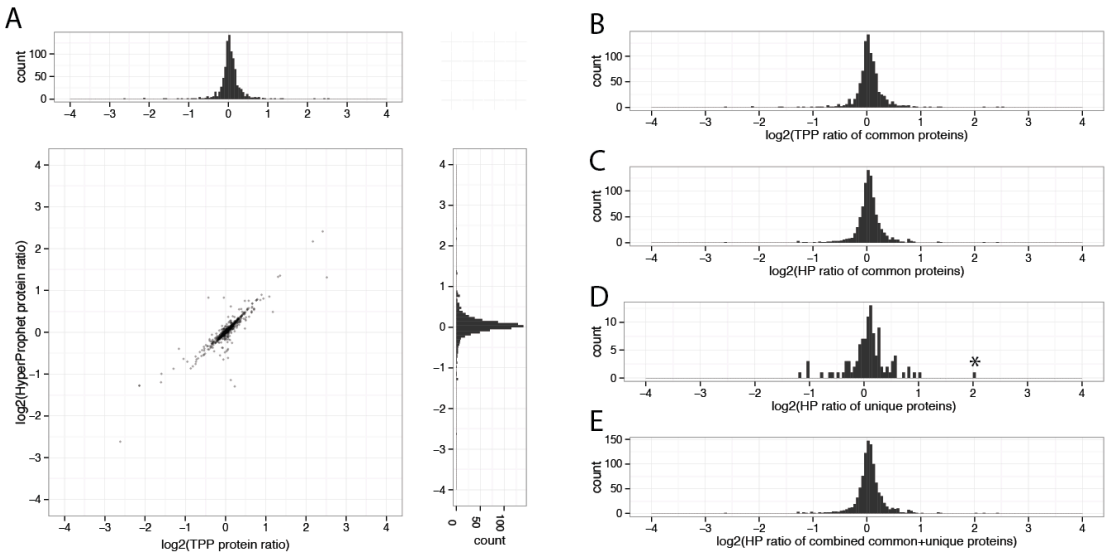


Figure 4.12 HyperProphet preserves SILAC protein ratios between replicates

In a test case comparing protein ratios before and after transferring peptides from a 1:1 SILAC tryptic yeast replicate analyses, protein ratios are preserved well as indicated by (A) a scatter plot of protein ratios either including or not including (TPP analysis) HyperProphet derived peptide quantitations. Note that only proteins in common can be plotted in this figure, therefore HyperProphet is contributing more peptides per protein and not new proteins. To assess the impact of HyperProphet on additional proteins, the distribution of protein ratios are re-plotted for (B) TPP protein ratios and (C) HyperProphet protein ratios, and also for (D) unique proteins only observed as result of HyperProphet, and (E) all proteins as a result of HyperProphet. The outlier labelled * is referred to in the text.

A distinctive feature of the new sample workflow proposed in this chapter is the use of single channel samples to aid in SILAC quantitation. Although a single channel sample may be derived from the same cell culture, its independent processing may result in variable tryptic digest efficiency and variable chromatography. Single channel samples should be considered as similar samples rather than replicates and it would be prudent to independently validate peptide transfers from single channel to SILAC samples. In the previous analysis between replicate SILAC data-sets, peptide ratios were compared between donor and recipient replicate analyses with scatter plots. In this analysis however, donor runs lack SILAC ratios. To observe the impact of HyperProphet on this experimental design, peptide ratios are plotted only as histograms to indicate the distribution of data (Figure 4.13). Again, transferred peptides result in ratios tightly clustered on 1:1 and have a similar level of precision as natively quantified peptides. At the protein level (Figure 4.14), an increase in the number of new proteins is observed, relative to transferring between SILAC analyses.

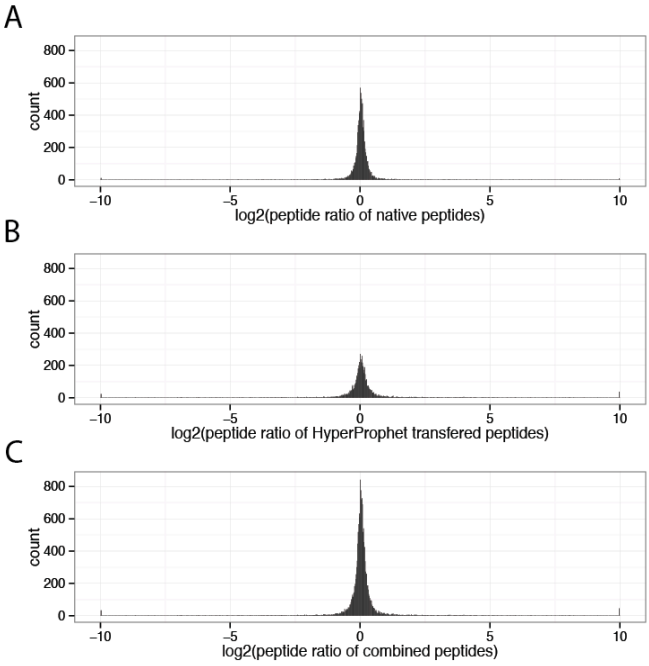


Figure 4.13 HyperProphet accurately transfers peptides from single channel analyses

Peptide ratio distributions for 1:1 SILAC tryptic yeast are clustered at 1:1 for (A) natively quantified peptides, (B) peptides transferred by HyperProphet from light and heavy single channel analyses, and (C) the combined set of peptides.

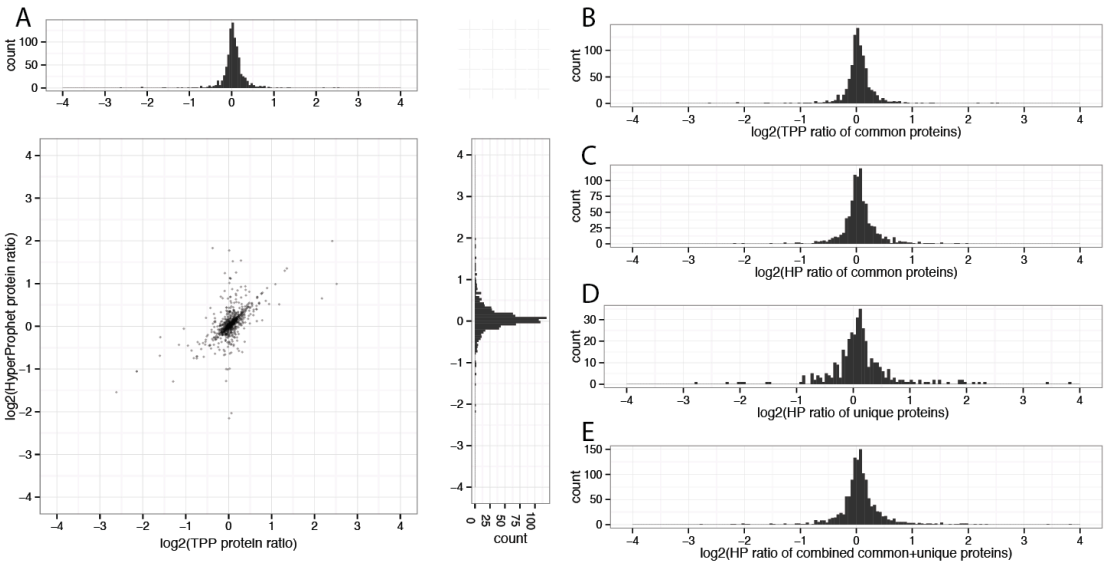


Figure 4.14 HyperProphet enables increased SILAC protein ratio determination using single channel data

In a test case comparing protein ratios before and after transferring peptides from a light and heavy single channel tryptic yeast replicate analyses, protein ratios are preserved well as indicated by (A) a scatter plot of protein ratios either including or not including HyperProphet derived peptide quantitations. The distribution of protein ratios are plotted for (B) TPP protein ratios and (C) HyperProphet protein ratios (including native and HyperProphet transferred peptide), (D) unique proteins only observed as result of HyperProphet, and (E) all proteins combined.

Given a potential for errors in HyperProphet derived quantitation, it is possible that quantitation of noise could tend towards a 1:1 distribution by chance. Although this is unlikely for high resolution instruments, to rule this out, analyses were repeated for SILAC samples mixed at approximately 2:1. Repeat analyses of figures 4.11-4.14 are presented for 2:1 SILAC samples in Figures 4.15-4.18. Results are equivalent for for 2:1 analyses indicating that quantitation of native and HyperProphet transferred peptides offer a high degree of precision. Total peptide numbers are slightly reduced in the 2:1 analysis compared to 1:1 data, although this is expected due to the reduced signal in the heavy channel. Also note that the data was not normalised, thus the ratio is slightly greater than 2:1 reflecting a small error in sample mixing.

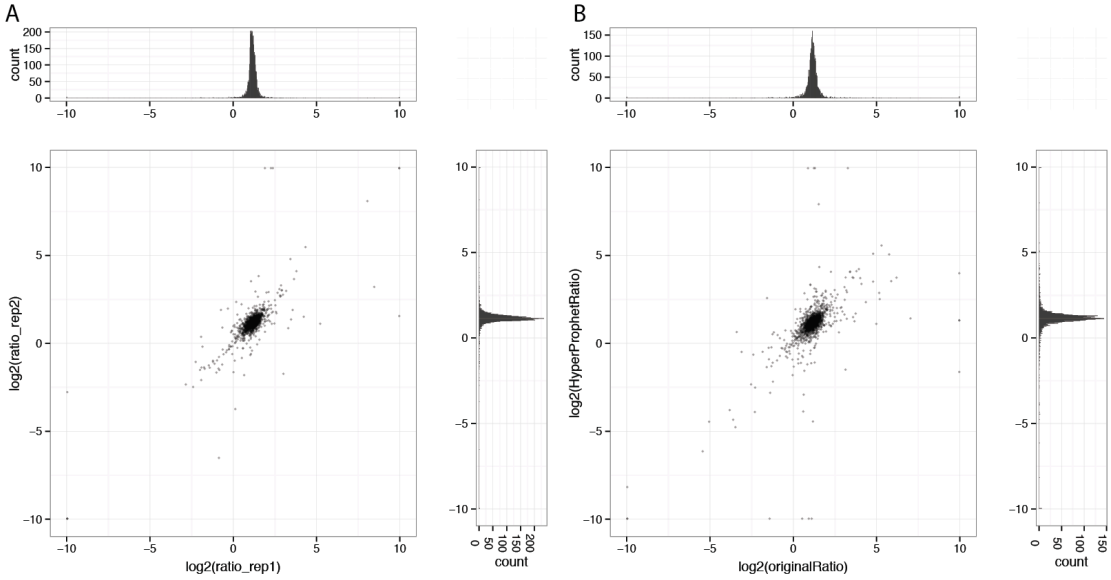


Figure 4.15 HyperProphet preserves SILAC peptide ratios between replicates at 2:1

In a test case between two 2:1 SILAC tryptic yeast replicate analyses, peptide ratios are well conserved between replicates as indicated by SILAC ratios for peptides in common between replicates (A). When using HyperProphet to transfer unique peptides (B), the ratio of transferred peptides in their recipient analysis (y axis) correlate well to the original SILAC peptide ratio in the donor analysis (x axis).

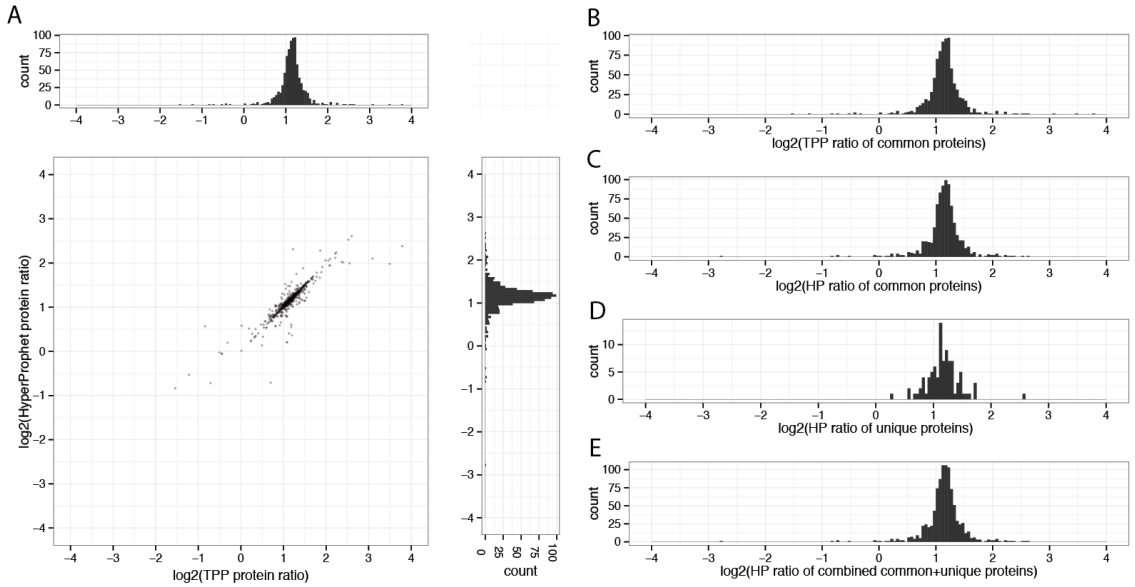


Figure 4.16 HyperProphet preserves SILAC protein ratios between replicates at 2:1

In a test case comparing protein ratios before and after transferring peptides from a 2:1 SILAC tryptic yeast replicate analyses, protein ratios are preserved well as indicated by (A) a scatter plot of protein ratios either including or not including HyperProphet derived peptide quantitations. To assess the impact of HyperProphet on additional proteins, the distribution of protein ratios are re-plotted for (B) TPP protein ratios and (C) HyperProphet protein ratios, and also for (D) unique proteins only observed as result of HyperProphet, and (E) all proteins as a result of HyperProphet.

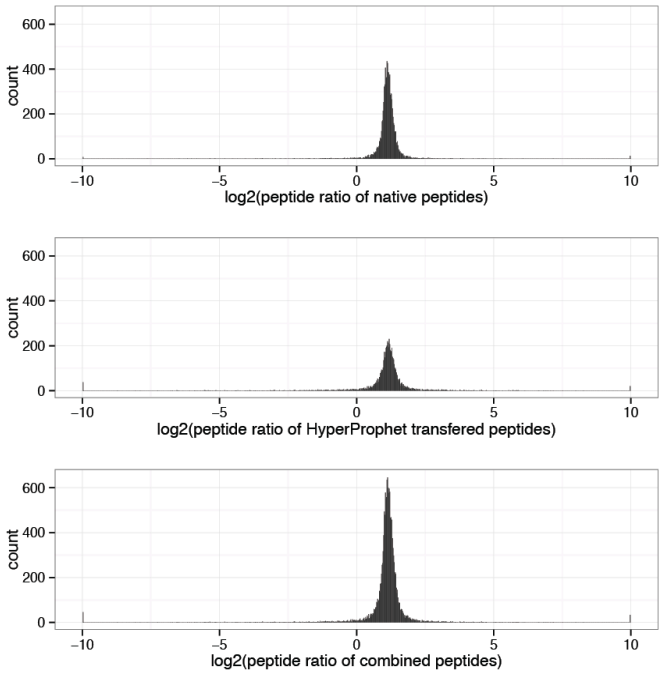


Figure 4.17 HyperProphet accurately transfers peptides from single channel analyses

Peptide ratio distributions for 2:1 SILAC tryptic yeast are clustered at 2:1 for (A) natively quantified peptides, (B) peptides transferred by HyperProphet from light and heavy single channel analyses, and (C) the combined set of peptides.

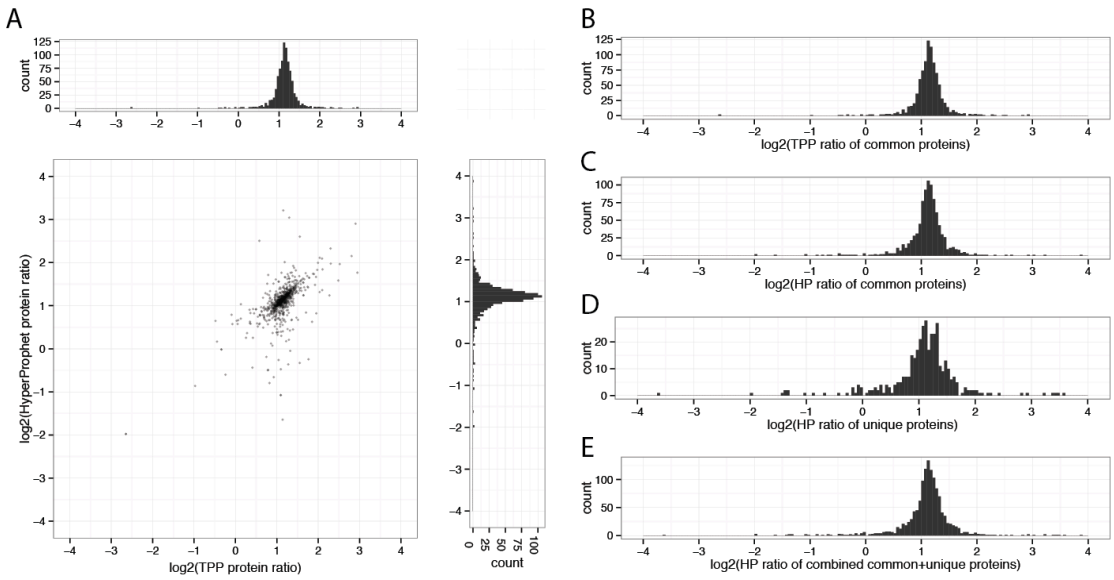


Figure 4.18 HyperProphet enables increased SILAC protein ratio determination using single channel data

In a test case comparing protein ratios before and after transferring peptides from a light and heavy single channel tryptic yeast replicate analyses, protein ratios are preserved well as indicated by (A) a scatter plot of protein ratios either including or not including HyperProphet derived peptide quantitations. The distribution of protein ratios are plotted for (B) TPP protein ratios and (C) HyperProphet protein ratios (including native and HyperProphet transferred peptide), (D) unique proteins only observed as result of HyperProphet, and (E) all proteins combined.

4.3.9 Managing false quantifications - SILAC label switch peptide filtering

During validation of HyperProphet, quantitative outliers were observed that affected both HyperProphet and normal TPP datasets. Extreme outliers may be false positive identifications with incorrect R/K compositions resulting in missing light or heavy SILAC partners. Interfering signals may also occur if two co-eluting peptides (or their isotopes) occur within the quantitative mass tolerance, a case of greater significance for non-fractionated and SILAC labelled samples where the complexity remains high. Whether HyperProphet is being used or not, the problem of how to deal with quantitative errors is an area of concern. A common approach to deal with unreliable quantitation is to use SILAC label switching where one or more replicates in an

experiment have their SILAC labels reversed. Subsequent analysis will require a pair of label switched replicates to have an inverse correlation between protein ratios (Butter et al., 2013), or more complex statistical analysis apply significance to a protein ratio over a set of label switched replicates (Ting et al., 2009). A protein showing a large deviation over the replicates due to label switching indicates unreliable quantification and will likely be rejected as being a significantly changing protein. This approach is similar to DNA microarray analyses, an academic field that has contributed to bioinformatics for proteomics (Gatto and Christoforou, 2014; Ting et al., 2009). Unlike microarrays, proteomics determines protein ratios from the assembly of peptide ratios rather than by direct protein detection. This fundamental difference allows us to approach SILAC label switching at the peptide level.

The concept for the peptide filter was published by (Lo et al., 2013) who presented a simple filter for dimethyl labelled bacterial analysis. In their approach, only two biological replicates were processed with opposite labelling orientation. Relative differences in ratios for peptides in common between analyses were calculated and the dataset ranked in order of relative difference. The most inconsistent ratios were discarded to preserve the peptides most conserved between label switched samples, resulting in an improvement of overall quantification results. However, their approach lacked the ability to filter more than two replicates, and an arbitrary 1.50-fold difference was used as a cut-off for up and down regulation, therefore neglecting downstream statistics. Also, an implementation was not supplied.

A label switch peptide filter was implemented in R (co-developed with supervisor, Dr. Patrick Pedrioli) to determine a subset of quantitatively reliable peptides over a multi-replicate SILAC TPP/HyperProphet analysis. Experimental design for the

peptide label switch is identical to the conventional protein level label switching, where SILAC replicates belong to one of two designs; sampleA(light)/sampleB(heavy) or label switched sampleB(light)/sampleA(heavy). Inverting design 2 *in silico* corrects the label switch so that the biological ratio is consistently expressed as sampleA/sampleB across all replicates. In a mathematically ideal scenario, each design will now correlate perfectly. This expected correlation is represented in Figure 4.19A as the dashed line. In a realistic sample, analytical variation will cause a deviation from the expected correlation, thus a small level of tolerance should be permitted (shaded region in Figure 4.19A). Erroneous peptide ratios will not be conserved after label switching and will fall outside the tolerance threshold and can be discarded as being quantitatively uninformative. For more advanced filtering, a non-linear tolerance is also permitted to account for greater variation with increasing ratios (Figure 4.19B), although a linear filter was used for research presented in this chapter. It should be noted that while peptides with large interferences or extreme false positive ratios are removed, genuine biological variation that is upheld by the peptide label switch is still maintained. It is inevitable that filtering peptides will result in a reduced proteome coverage, particularly due to the requirement for peptide ratios to be observed in each design. In contrast, HyperProphet increases proteome coverage making HyperProphet and the SILAC label switch peptide filter an ideal partnership.

A high-level description of the SILAC label switch peptide filter function follows:

- Take as input
  - a set of quantified peptides for each replicate
  - a design matrix (1 or -1 for each replicate)
  - A quadratic formula describing tolerance
- Median normalise each set of peptide ratios
- Divide replicates into design 1 (A/B) and design -1 (B/A)

- For each design, determine set of peptide ratio groups (all peptides representing the same ratio, i.e. a light and heavy peptide can represent the same ratio)
- Determine mean ratio for each peptide ratio group, inverting ratio for design -1
- Pair corresponding peptide ratio groups between designs, discarding non-paired peptides (data points on scatter plot in figure 4.20A)
- For each data point, determine length of orthogonal projection on the expected correlation line and calculate permitted tolerance, t
- Determine perpendicular distance, d, from expected correlation line
- Delete peptide ratio group if d > t
- Filter original input data such that peptides belong to remaining peptide ratio groups
- Recalculate protein ratios (geometric mean) using remaining quantitatively reliable peptides

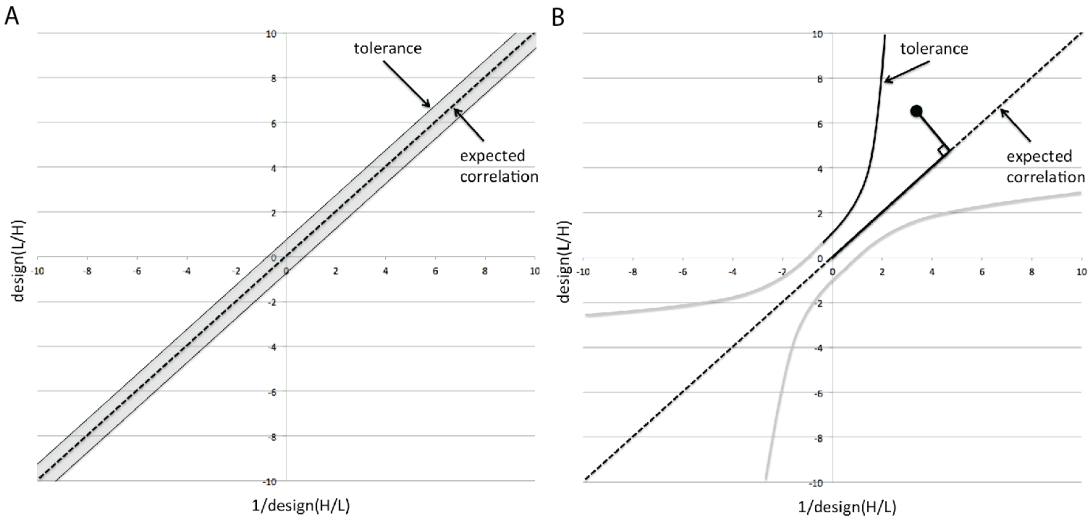


Figure 4.19 Conceptual design of SILAC peptide label switch filtering.

After inversion of SILAC label switched peptide ratios, accurate log2 ratios from each design are expected to correlate on x=y (dotted line). A linear (A) or a polynomial (B) function can be applied to define a subset of peptides ratios within a tolerated range. The tolerance is a function of the orthogonal projection length on the expected fit line, and a point is tolerated if the perpendicular distance from the line is less than the tolerance. As all distances are positive, the user defined function is applied in the positive quadrant (black line) and is symmetric in all 4 quadrants.

To demonstrate the effectiveness of the filter, the filter was applied to 6 biological LCMS replicates of *elp3*Δ yeast (this data will also be used to exemplify HyperProphet in the next section). Three replicates were acquired as *elp3*Δ/wt and 3 further with label switching to wt/*elp3*Δ. All data was process through TPP without the use of HyperProphet. Figure 4.20 provides a graphical overview of the filtering, and is the actual output generated when executing the filter. Pre-filtered peptide ratio groups in Figure 4.20A show a dense cluster at 1:1 with a distinct trend at x=y over a range of approximately +/- 2-fold. However, the greatest range of data points is actually at x=-y and extends to each extreme ratio from -10/10 and 10/-10. These extreme ratios are no longer present after filtering (Figure 4.20B) and the maximal peptide ratios group means are now approximately +/-2 (4-fold in normal space). The impact the filter has on peptide ratio distribution for each replicate is displayed in Figure 4.20C-D. The box plots show that the majority of the extreme data points have been removed by the filter.

Many proteomic studies have an interest in the quantitation of individual peptides, particularly those investigating post-translational modification such as ubiquitylation and phosphorylation. In such cases, this filter can offer improved reliability of individual peptide ratios. For proteomic studies focused on protein quantitation, multiple peptides per protein are typically required for quantitation. To investigate the impact on protein ratio calculation, the mean and standard deviation was calculated for each protein using peptides assigned to it by TPP. Surprisingly, the distribution of means is only slightly improved after filtering (Fig 4.21 A and B) suggesting that peptide errors are being averaged out for proteins with multiple peptide assignments. However, the standard deviations are drastically reduced (Figure 4.21C-D) despite the removal of peptide by the filtering process (note that reducing a sample size will increase a standard deviation). Increased precision at the peptide level improves protein

ratio reliability, which is of particular importance for proteins with only a few assigned peptides available for quantitation.

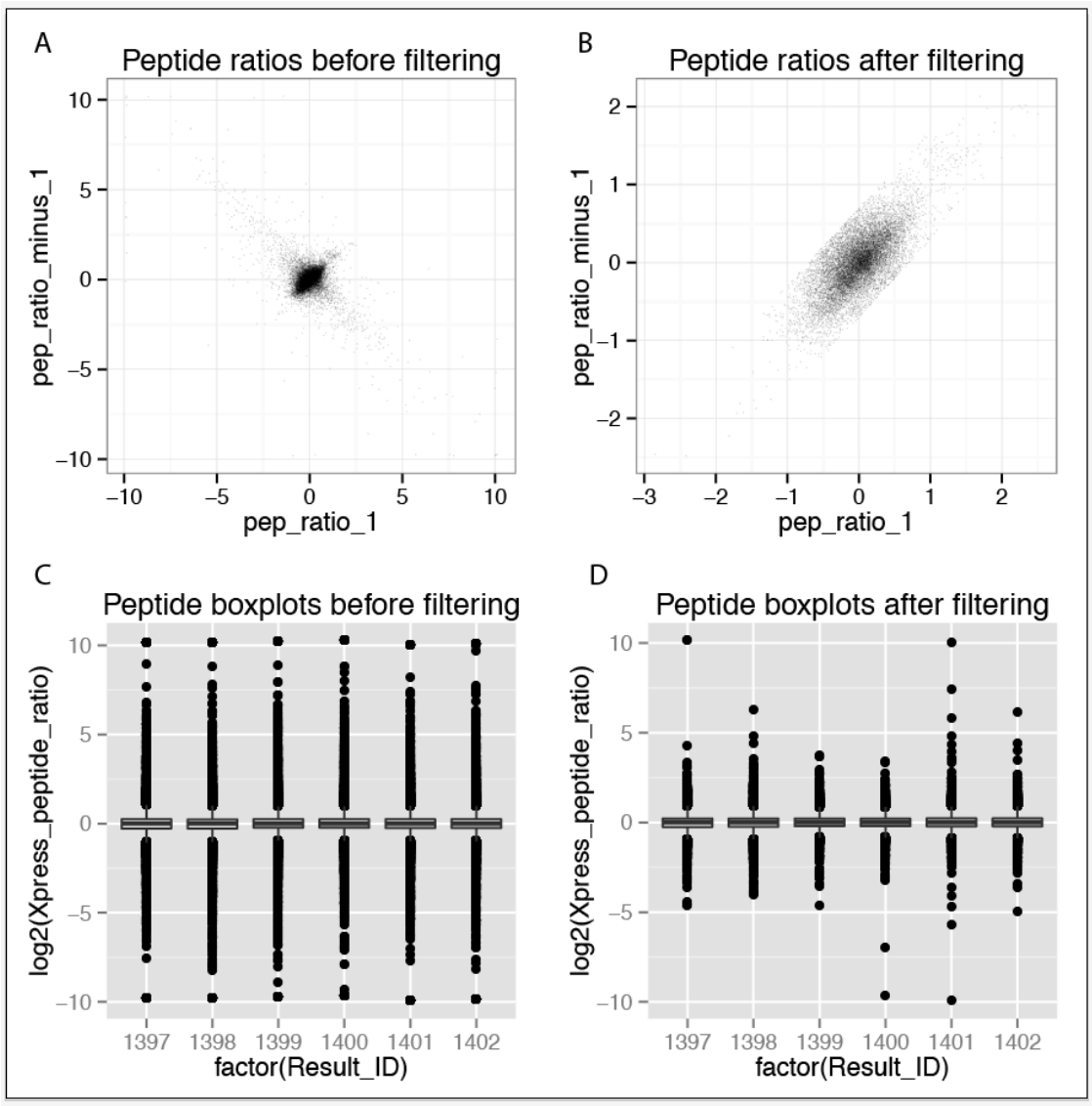


Figure 4.20 Demonstration of SILAC label switch peptide filter

The SILAC label switch filter (linear +/-0.5 tolerance) has been applied to 3 biological replicates each of *elp3*Δ/wt and wt/*elp3*Δ yeast. Log2-fold ratios for means of peptide ratio groups within design 1 (x-axis) and design -1 (y-axis) are displayed (A) before and (B) after filtering. Distribution of peptide ratios for each replicated (C) before and (D) after applying filter demonstrates removal of many extreme outliers in peptide datasets.

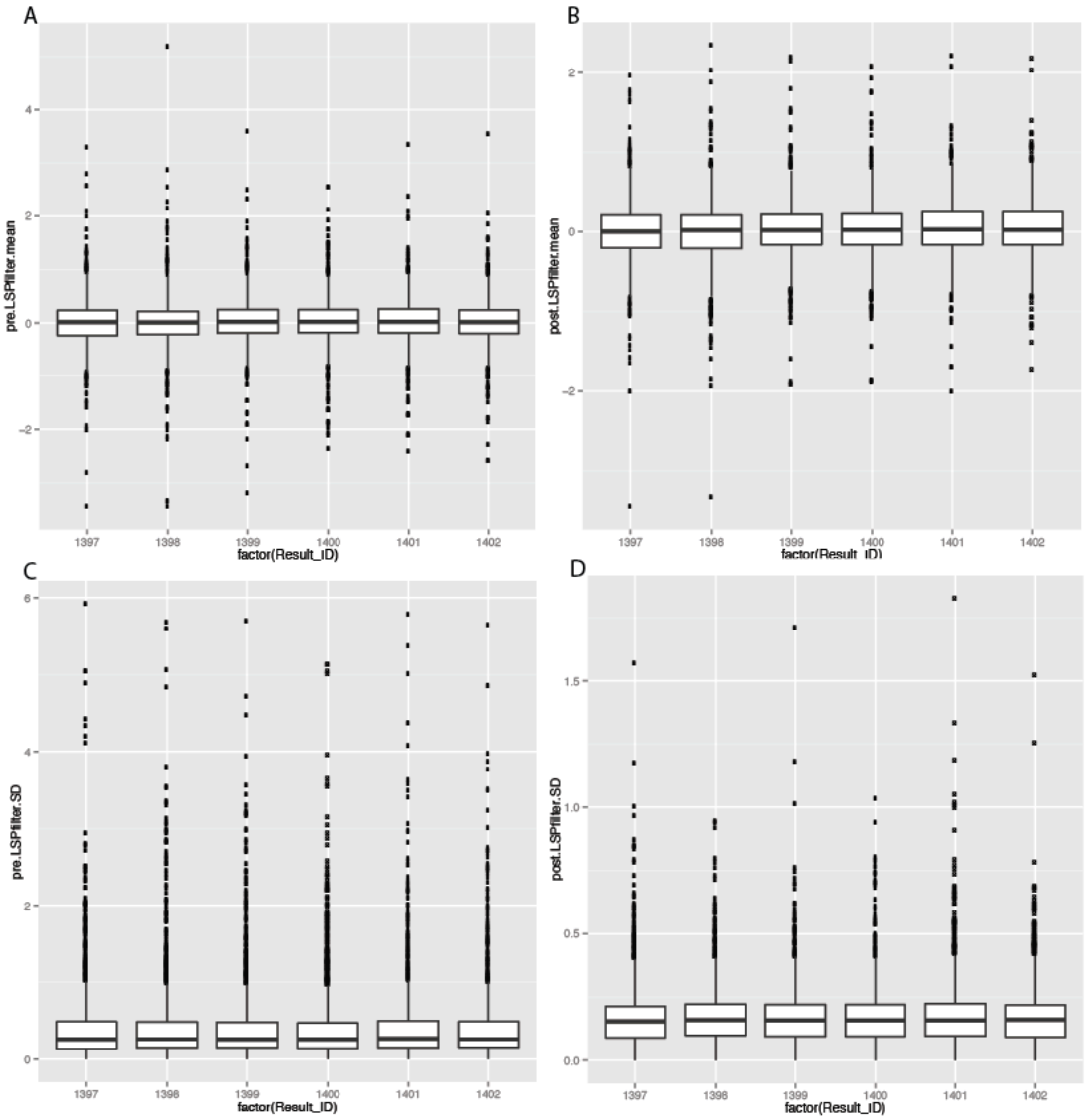


Figure 4.21 SILAC label switch peptide filtering improves precision of protein ratios.

Protein-centric peptide mean and standard deviations are presented for 6 replicates of *elp3*Δ. The distribution of means before (A) and after (B) filtering is slightly improved. Standard deviations before (C) and after (D) filtering indicate a striking improvement in protein ratio precision.

The greatest impact of improved precision is observed during statistical analysis of protein ratios over replicates. Consistent with previous observations for individual proteins, the distribution of means for protein ratios over 6 replicates is slightly reduced but has vastly improved standard deviations (Figure 4.22). The importance of the reduced deviation becomes evident when considering the purpose of applying a t-test to

the SILAC data; to determine if the set of SILAC ratios are significantly different from 1:1, based on their mean and standard deviations. Furthermore, an improved population error can further impact a Bayes t-test, as is used in this chapter, because error estimates are adjusted towards a population estimate.

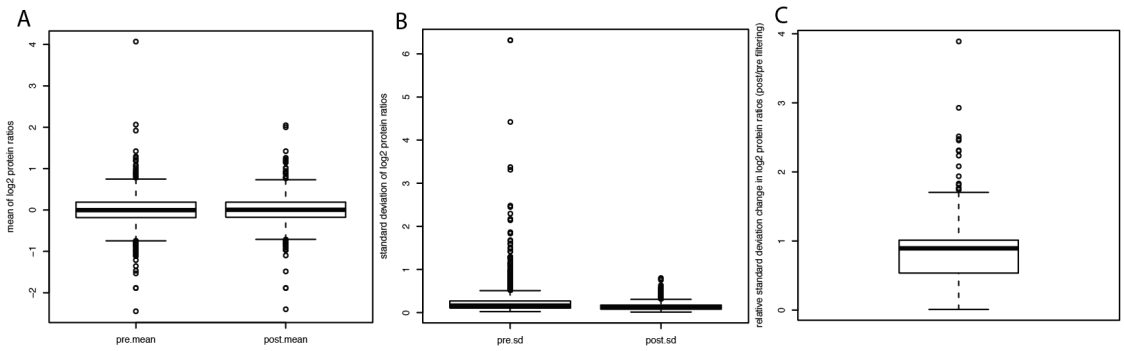


Figure 4.22 SILAC label switch peptide filtering improves experiment-wide precision in protein ratios.

Protein means (A) and standard deviations (B) over 6 replicates of *elp3*Δ are presented pre-filter (left) and post-filter (right). The distribution of protein means is slightly improved, whereas standard deviations show a striking improvement. A relative change in standard deviation (B), where pre-filtered standard deviations are defined as 1, indicate that deviations have reduced for approximately 75% of the data.

Figure 4.23 compares the result of a Bayes moderated t-test on the 6 replicates before and after SILAC label switch peptide filtering. The reduced variation is evident in this graphical representation, particularly for non-significant proteins at low probability, and has resulted in increased number of significant protein changes with detection of smaller fold-changes. Contrary to conventional protein level SILAC label switch filtering, which discards proteins with high variance, the peptide level filtering has enabled us to prevent needless rejection of proteins due to anomalous peptide quantifications.

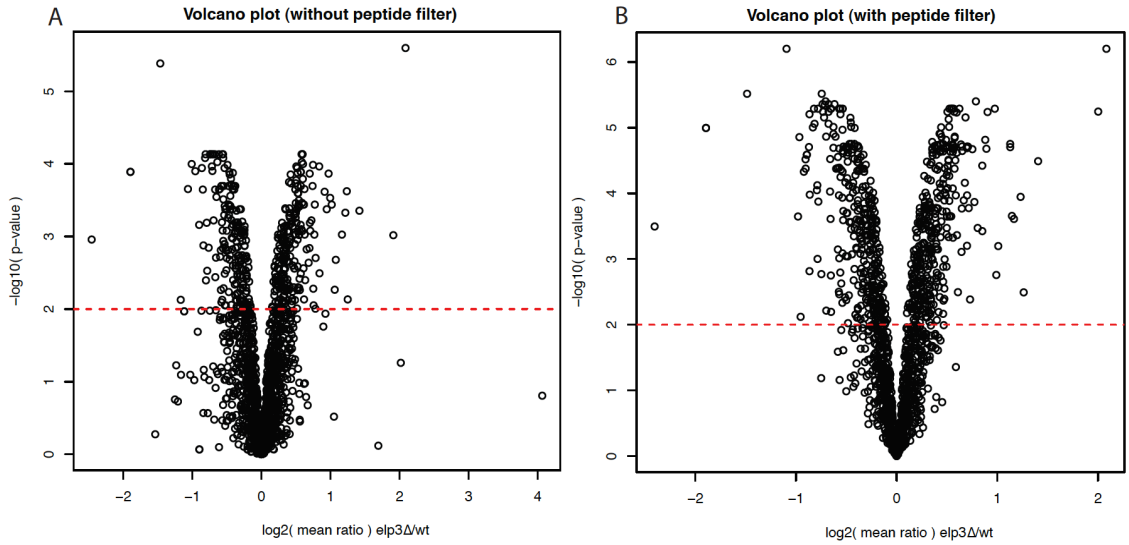


Figure 4.23 The effect of SILAC label switch peptide filtering on a Bayes moderated t-test

Data from 6 biological replicates of *elp3*Δ yeast were subjected to SILAC label switch peptide filtering and the resulting protein ratios were then subjected to a Bayes moderated t-test. Volcano plots of log2 fold change against FDR-adjusted significance are presented for data (A) before and (B) after filtering peptides. The red line indicates a 1% false discovery rate.

4.3.10 Impact of experimental design on proteome coverage

Various experimental designs have been investigated to assess the impact of using the combination of the SILAC label switch peptide filter, transferring peptide identifications with HyperProphet, and the use of single channel samples. Six biological LCMS replicates of *elp3*Δ/wt yeast were acquired, 3 replicates *elp3*Δ/wt and 3 label switched to wt/*elp3*Δ (as was used to exemplify the label switch filter in the previous section). Two single channel samples were also sampled at the time of mixing SILAC samples, wt light (unlabelled) and wt heavy (fully labelled). Four different experimental designs were composed:

TPP6(no filter); 6 SILAC replicates, using a 'typical' TPP workflow without additional features.

TPP6; 6 SILAC replicates, using label switch peptide filter (but not HyperProphet).

HP6; 6 SILAC replicates, using label switch peptide filter and HyperProphet.

HP4+2; 4 SILAC replicates and 2 single channels, using label switch filter and HyperProphet.

Note that each experimental design builds upon the previous starting with a TPP 'first pass' dataset, and then the addition of peptide filtering, HyperProphet, and single channels, respectively. The label switch peptide filter, where used, was applied with a linear +/-0.5 tolerance and HyperProphet was executed with identical parameters other than the altered choise of input input files. Also note that the last experimental design maintains a total of 6 LCMS analyses for a fair comparison to another designs. In addition to using a 1% FDR peptide and protein threshold, only quantified peptides were used, proteins were also required to have at least 2 quantified peptides, and proteins must be observed in at least 4 replicates. On average, approximately 20% of identified proteins did not pass the filter, however, this rigorous filtering permits comparisons to be made strictly from a quantitative perspective.

Figure 4.24 presents peptides and proteins quantified for each experimental design. Label switch filtering peptides has an expected negative impact on proteome coverage due to discarding peptides. The reduction in quantifiable peptides can be observed at the peptide level for both TPP and HyperProphet analyses (Figure 4.24A). Likewise, a reduction in quantified proteins is observed, however, the loss of proteome coverage is more than compensated for by applying HyperProphet to the dataset (Figure 4.24B). A further increase in proteome coverage is observed with the inclusion of single channel analyses. Considering that the final experimental design replaces SILAC samples with single channel samples, rather than additional analyses, this boost in proteome coverage is clearly a consequence of the supplemental proteome coverage offered by single

channel samples. The use of HyperProphet and single channels has boosted the proteome coverage from 1559 to 2112 quantified proteins, which represent a 35% gain.

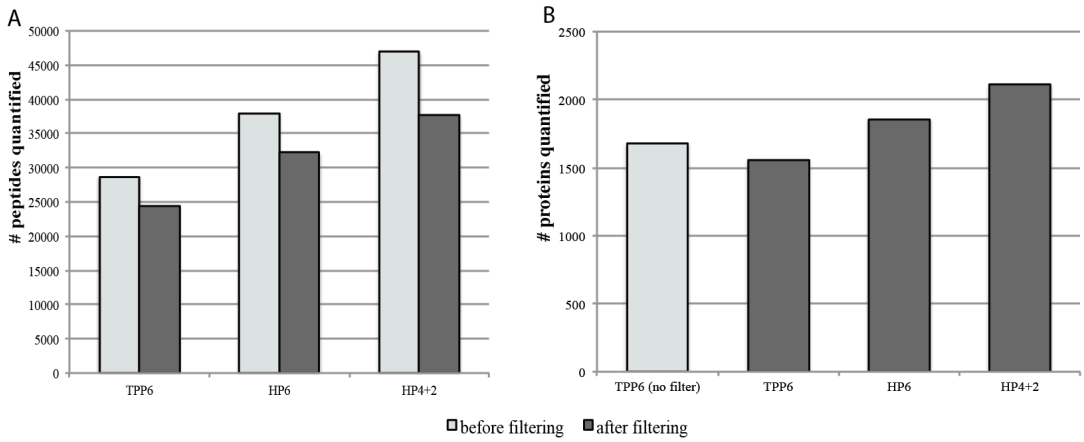


Figure 4.24 HyperProphet and the use of single channel analyses increase peptide and protein coverage

(A) Total peptides quantified before (light bars) and after (dark bars) label switch peptide filtering demonstrate the advantage HyperProphet has over just TPP when transferring peptides from SILAC and single channel analyses. An similar advantage is demonstrated for quantified proteins (B). Note particularly that HyperProphet analyses exceed the proteome coverage of a traditional TPP analysis even after filtering out quantitative outliers.

4.3.11 Impact of experimental design on statistical significance

While maximising proteome coverage is important for discovery proteomics, the power of a statistical analysis is also an important consideration. Statistical power is the ability to detect significant protein changes of a given magnitude, and is function of sample size and measurement precision (Levin, 2011). It follows that increasing sampling precision or size will increase the confidence in true protein changes (i.e. smaller p-value) or enable detection of smaller changes in the proteome. Improvement of precision has already been demonstrated using the SILAC label switch peptide filter. The number of replicates in which a protein has been observed can also increase

statistical power. To this effect, not only does HyperProphet increase proteome coverage, but it also reduces null values across replicates. The incidence of 6 out of 6 observations increased from 88% to 95% between TPP and HyperProphet. This effect would also have a positive impact for analyses where null values are particularly undesirable, such as time course proteomics.

For a numerical assessment of statistics analysis over replicates, a Bayes moderated t-test was conducted for each experimental design. The resulting total number of proteins significantly changing in each test is compared in Figure 4.25A. Results are presented for both 1% FDR and 5% FDR as these are two thresholds commonly used in the literature. The label switch peptide filter increases the number of significantly changing proteins considerably in a TPP analysis despite a slight reduction in protein coverage (data reproduced from Figure 4.23). By applying HyperProphet, the detected number of proteins changing increase further and proportionately with the increase in quantified proteins. The combination of HyperProphet and label filter have almost doubled the number of high confidence 1% FDR changing proteins. In contrast, the experimental design including single channel samples results in a drop in detected protein changes despite improving proteome coverage. This is an intuitive result when considering that the number of quantifiable replicates has reduced from six to four. In addition to a drop in quantitative sensitivity, there is also a noticeable difference between 1% FDR and 5% FDR. This is also a consequence of the reduced sample size and can be observed in Figure 4.25 B-C as a compression in the higher probability set of proteins. The benefit of this new workflow becomes even more apparent when comparing the last experimental design to a typical TPP workflow without HyperProphet or the label switch filter. Even with only four quantifiable replicates, there are more significantly changing proteins detected than with 6 replicates without

using these tools. The ability to maintain or increase quantitative sensitivity while reducing replicates offers an advantage to studies where sample preparation procedures are difficult or expensive.

On designing an experiment, it is clearly advantageous to acquire as many quantifiable replicates as is practical to maximise statistical power. The inclusion of single channel samples clearly gives an advantage in proteome coverages. However, these results demonstrate the importance that single channel analyses do not replace quantifiable SILAC analyses to avoid loss of statistical power.

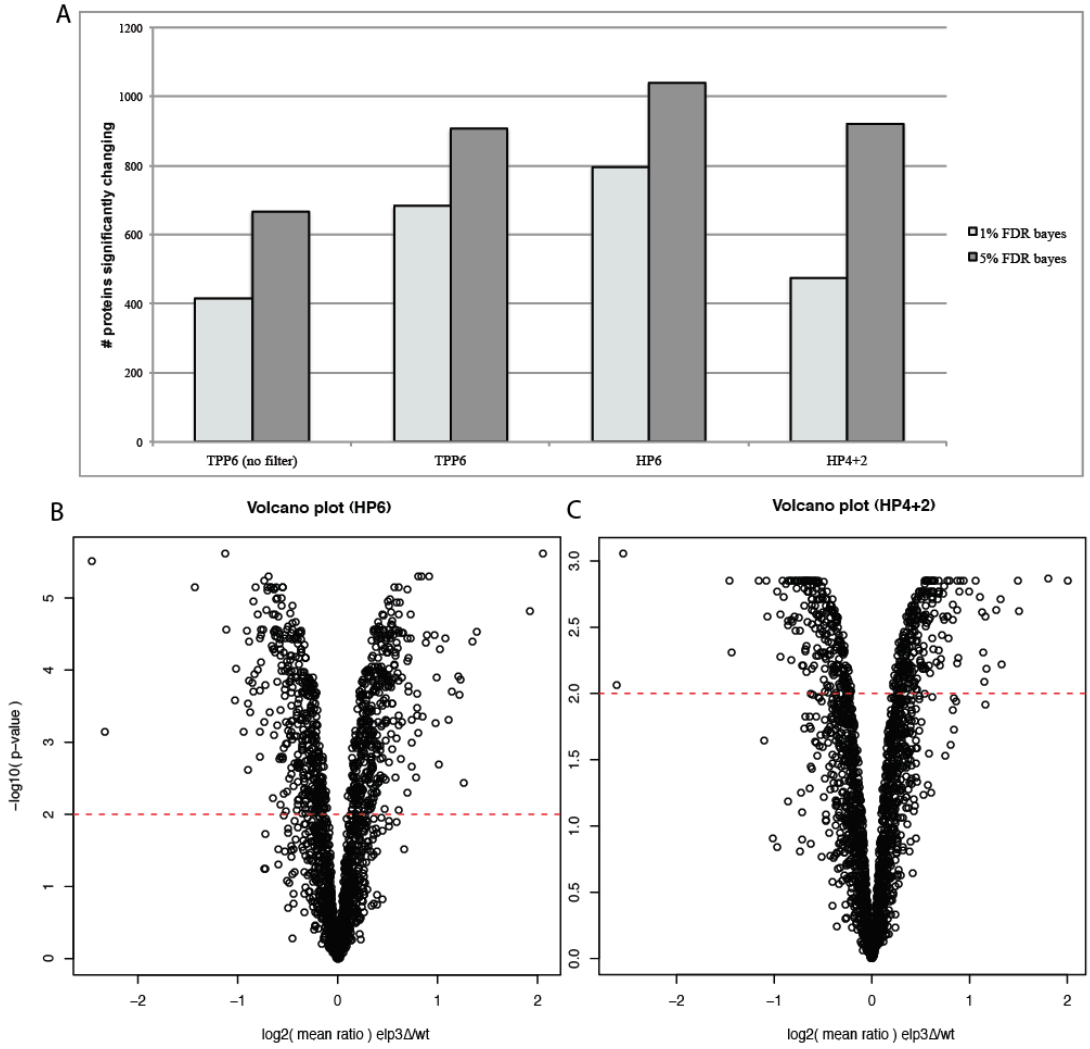


Figure 4.25 Impact of experimental design on statistical power

Total proteins detected as significantly changing at 1% and 5% FDR in a Bayes moderated t-test with p-value correction (A) indicate that each design has a large impact on the statistical test's proteome depth. The impact of replacing SILAC analysis with single channels is displayed as volcano plots for HyperProphet analyses of 6 SILAC

replicates (B) and 4 SILAC replicates with 2 single channel analyses (C). Proteins above the red line are significantly changing at 1% FDR.

4.3.12 Codon bias analysis of *elp3*Δ

Deletion of *ELP3* results in hypo-modification of tRNAs and reduced translation efficiency, therefore a widespread effect on the proteome is expected. However, the volcano plots in Figure 4.25 indicate that there are very few proteins with large expression changes and the effect of *ELP3* deletion on the proteome is subtle. As a result, we require a sensitive analysis to penetrate as deep into the proteome as possible to detect small protein changes. The HyperProphet analysis using 6 SILAC replicates with a 1% FDR threshold was selected for further analysis due to its strongest statistical power.

A random forest analysis of gene codon composition was conducted to classify proteins as up or down regulated. Figure 4.26 displays the ranking of each codon, with AAA, CAA, AGA, GAA, and AAG being identified as the most important in predicting significant changes in protein abundance. To control for codon bias analysis reproducibility, the analysis was repeated 10 times with 10 random seeds (Figure 4.27A). The variable importance axis is unit-less and should not necessarily be comparable between analyses, thus standard deviations will potentially overestimate variability between random forest results. However, results were found to be highly reproducible, with the top six codons remaining in the same order. Overlapping error bars indicate that the exact order of lower ranking codons should not be given too much weight. A repeat analysis using randomised data (Figure 4.27B) failed to reproduce the findings. A few codons were reported above the significance line suggesting a small amount of over-fitting can occur, so only codons reported well above the significance line should be considered as biologically relevant. The replicate and randomised

controls give support the codon bias results as robust, and also supports the proteomics data quality, which is a valuable validation since we are discussing small fold-changes.

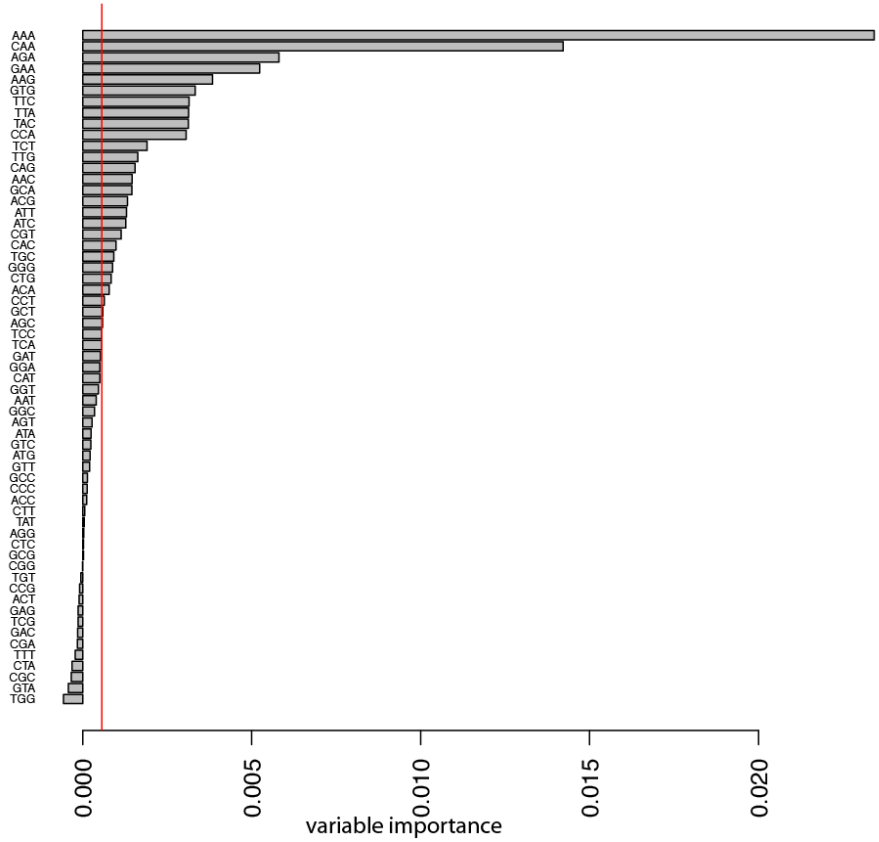


Figure 4.26 Codon bias analysis of the top 1% FDR proteins changing in *elp3*Δ/wt yeast.

A random forest algorithm determines the gene frequency of AAA, CAA, AGA, GAA and AAG to be most important for predicting the up or down expression of proteins in *elp3*Δ.

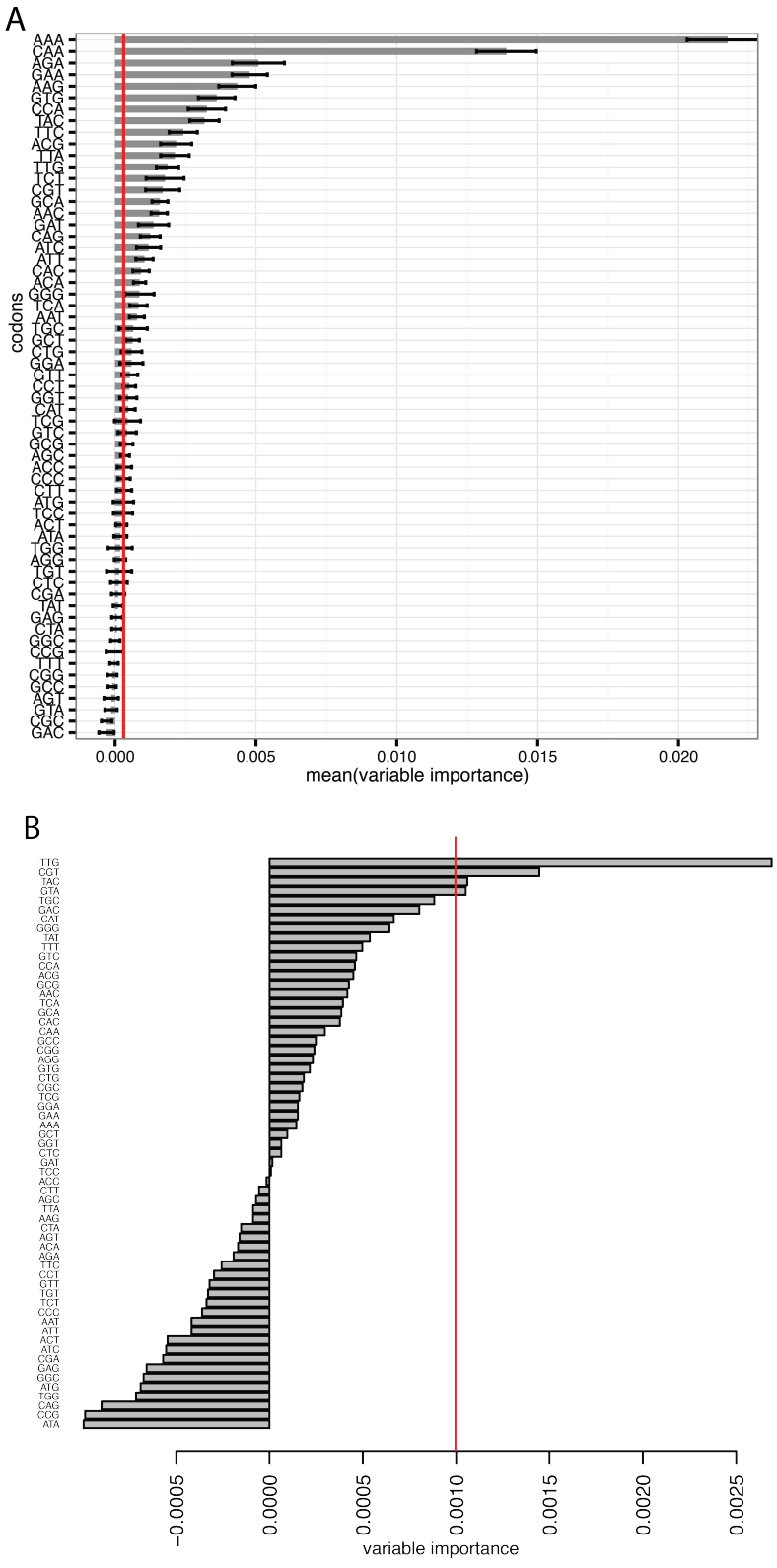


Figure 4.27 Replicate and randomised analyses give confidence to *elp3*Δ codon bias analysis.

(A) Codon bias analyses (as figure 4.26) mean and standard deviations over 10 repeat analyses using 10 random seeds selected between 10-99. (B) A repeat analysis where up or down regulation has been randomised over the same set of proteins in A.

4.3.13 *elp3*Δ codon bias overlaps with *urm1*Δ

It is a notable observation that the most biased codons include AAA, CAA and GAA, which are the three codons recognised by the doubly modified mcm5S2U tRNAs. In a similar codon analysis of *urm1*Δ, the sulphur carrier responsible for thiolating tRNA wobble uridines for tRNA-K(UUU), tRNA-Q(UUG), and tRNA-E(UUC), strong codon biases were reported for AAA, CAA, AAG, GGG, GAA (Rezgui et al., 2013). A direct comparison to this analysis reveals that, in addition to the three thiolated tRNA cognate codons, AAG is also in common. This overlap suggests that the primary effect of deleting *ELP3* is strongly related to its involvement in the co-modification of uridines with the *URM1* pathway. Observing some overlap is unsurprising given that aberrations in these two pathways have similar phenotypes, and there is in-vitro and in-vivo evidence that both *URM1* and *ELP3* are required for efficient translation (Rezgui et al., 2013).

The AAG codon bias is also of interest as this was also observed in the *urm1*Δ analysis. AAG, which encodes for lysine, differs from the AAA lysine codon only at the wobble position and AAA was the most biased codon. This suggests that AAG is likely to be involved with the modification state of the mcm5S2U modified tRNAs, despite the fact that the tRNA recognising AAG is not modified at the wobble position. It is unclear why this codon would present a bias under these conditions. Murphy et al. (2004) reports that tRNA-lys(UUU), which binds its cognate codon AAA, can also bind its near-cognate AAG codon but only if modified. The codon bias could be explained by the tRNA wobble from AAA to AAG codons being an important mechanism for efficient translation of lysine coded by AAG. An alternative hypothesis could be that wobble can occur without modification, contradicting Murphy et al. (2004), but with reduced efficiency. Poor binding to AAG could therefore cause a translational defect

for genes rich in this codon. Over-expression of the unmodified tRNA-lys(UUU) can compensate for the lack of modification and ameliorate symptoms of *elp3Δ* (Esberg et al., 2006), despite lacking the modification that aids recognition of the near-cognate codon. It would be interesting to see if the AAG bias remains with over expression of tRNA-lys(UUU).

As the *elp3Δ* codon bias was revealed to have a striking similarity to *urm1Δ*, it is worth considering the cause of the shared phenotypes. It is clear that both *URM1* and *ELP3* are required to form the doubly modified mcm5s2U (Huang et al., 2008), and the presence of at least one modification appears to be essential as double mutants are lethal (Costanzo et al., 2010). The similarities between *elp3*Δ and *urm1*Δ phenotypes could occur because both tRNA modifications are equally required and removal of either mcm5 or s2 impairs translation efficiency. However, the relationship between these modifications is complicated by the fact that disruption of the Elp-complex also results in reduced thiolation (Leidel et al., 2009).

4.3.14 *elp3*Δ specific codon bias response

Of the eleven tRNAs that receive ELP-dependent modifications, eight are independent from the doubly modified mcm5S2U tRNAs. AGA is also ranked highly in the codon bias analysis, which is modified by mcm5U and is the most heavily used of the 6 arginine codons (Iben and Maraia, 2012). This codon was not reported in the *urm1Δ* codon bias analysis suggesting that this is a result specifically due to the absence of mcm5 modification on tRNA-Arg (UCU). This observation is aligned well with reports of TRM9 deficient cells, which are unable to methylate cm5U to mcm5U, and have impaired translation of AGA codons (Begley et al., 2007). This would suggest

that it is specifically the methyl modification that is required but either the loss of methylation or complete loss of the modification has a similar effect for this codon. There have also been reports that the UCU codon can be thiolated in the bovine liver (Keith, 1984). It could be that tRNA-Arg (UCU), like the mcm5S2U modified tRNAs, has a larger dependence on modifications than most tRNAs for efficient translation and higher eukaryotes have a developed additional mechanisms to modulate translational efficiency.

Given that the yeast strain being used is arginine auxotroph, this yeast strain may be very sensitive to arginine availability. An alternative explanation for the AGA codon bias could be a reduced intracellular arginine availability which in turn is limiting availability of charged arginine tRNAs. This would be manifested by an AGA bias, as this is one of six arginine codons but represent almost 50% of the arginine codon usage (Iben and Maraia, 2012). Although this appears to be an ELP specific effect, there is no evidence to suggest whether this is a direct effect of the tRNA modification or an indirect transcriptional effect altering arginine availability. It would be prudent to validate this in an arginine prototroph, which can still be done via SILAC proteomics but using only lysine and lysC-endopeptidase.

The remaining codons corresponding to ELP-modified tRNAs did not show a noticeable bias over other non-modified codons. ELP-dependent tRNA modifications therefore have negligible impact on translation compared to the dominant codons already discussed. This does not rule out the existence of a translational defect for other codons. It is likely that, if a bias exists for other codons, they were not clearly distinguishable by the method used due to a dominant effect from the codons normally recognised by mcm5S2U-tRNAs. To uncover other potential subtle effects on the

proteome, it would be necessary to isolate the effect of ELP-mediated modifications from effects of the three doubly modified mcm5s2 tRNAs. This might be achievable by over expression of tRNA-K(UUU), tRNA-Q(UUG), and tRNA-E(UUC), which has been shown to suppress phenotypes of both *elp3Δ* and *urm1Δ*. A direct comparison between *elp3Δ* and *urm1Δ* would be the most straight forward approach to remove the common codon bias effects. As previously discussed, a *elp3Δ*/*urm1Δ* SILAC comparison would be a poor way to differentiate between the mcm5 and s2 modifications on doubly modified tRNAs due to the interdependence of the modification pathways. It would however provide an opportunity to study ELP-specific effects, including codon biases associated with ELP-dependent modifications on non-thiolated tRNAs.

4.3.15 *elp3*Δ proteome analysis and comparison to *urm1*Δ

The quantitative proteome analysis revealed 398 up regulated proteins and 398 down regulated proteins at 1% FDR. Very few proteins were observed to change at any considerable amount though. Only 20 proteins were found to be changing by more than 2-fold. A gene ontology enrichment (GO) analysis was conducted on the 1% FDR up and down regulated proteins using FunSpec (Robinson et al., 2002). Down regulated proteins were enriched for anabolic processes such as translation and glycolysis. Up regulated processes were enriched for catabolic processes such as proteasomal degradation, and also amino acid biosynthesis (the full list of proteins and GO results are in Appendix E). The most significant functional and biological processes are in common with the *urm1Δ* GO analysis publish by Rezgui et al. (2013) indicating that both *elp3Δ* and *urm1Δ* have very similar effects on the proteome. Furthermore, of the six proteins confirmed to be down regulated in *urm1Δ* cells by western blotting, five of

them were also found to be down regulated in *elp3Δ* (CMS1, YPL199C, FPR4, DEF1, BFR1). The sixth, mcm1, was not quantified in this study.

It is tempting to speculate upon the biological significance of the protein changing in the *elp3Δ*/wt analysis. However, given that *elp3Δ*/wt and *urm1Δ*/wt proteome analyses were found to be highly similar in both the codon bias and GO analyses, it follows that the majority of the significant proteome changes are in fact a result of codon bias due the convergence of each pathway on the same set of tRNAs. In a direct comparison of *elp3Δ*/*urm1Δ* SILAC by Rezgui and colleagues, proteomes were found to be very similar and only 12 proteins were significantly changing. To determine if this observation is supported by this study, *elp3Δ*/wt was compared to the *urm1Δ*/wt data extracted from Rezgui et al. (2013). Figure 4.28 displays a scatter plot of proteins in common between each analysis. The correlation is surprisingly strong (R2 = 0.51) given that they are of different mutants, the datasets were collected by different means (fractionation versus non-fractionated), and the analysis workflow differed substantially. The vast majority of proteins significantly changing are also changing in the same direction in the other mutant, indicating that *ELP3* and *URM1* have very similar overall effects on protein expression.

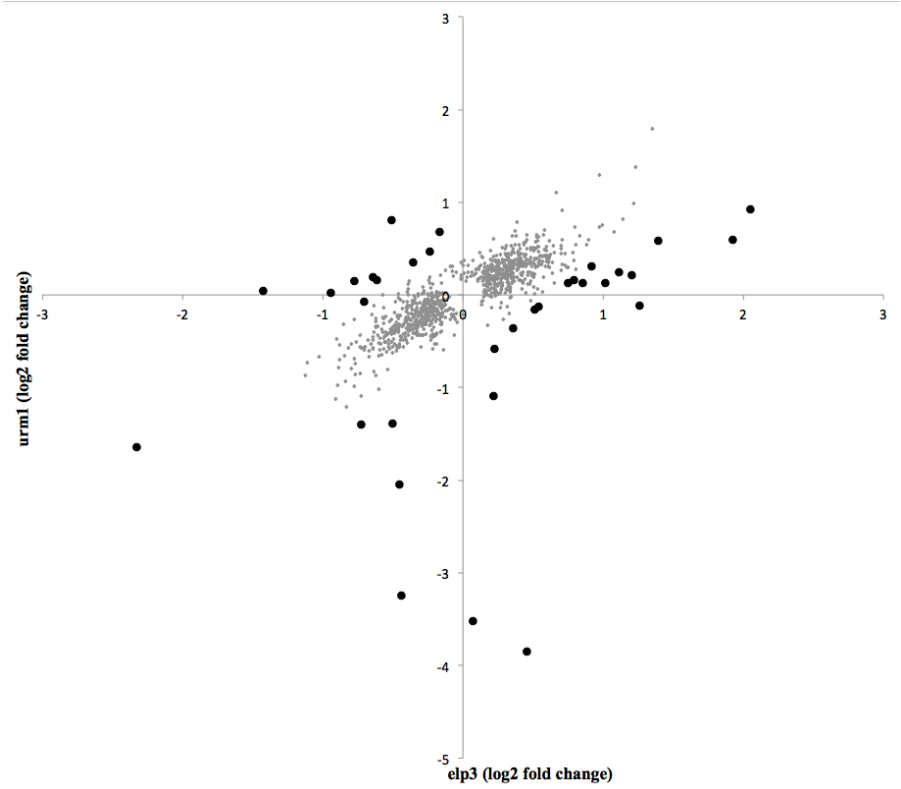


Figure 4.28 Comparison of *elp3Δ*/wt and *urm1Δ*/wt proteomes

A scatter plot of protein log2 fold change for proteins in common between datasets from *elp3Δ* (this study) and *urm1Δ* (Rezgui). Differences >1.5 fold between datasets are in bold and are listed in table 4.1.

There are however obvious differences outside the main cluster of conserved proteins. To identify obvious differences between datasets, an arbitrary cut-off of 1.5-fold was applied. To minimise false positive discoveries between datasets, additional filtering was applied such that all proteins compared were significantly changing in at least one of the studies (resulting proteins are listed in Table 4.2). REP2, YGP, ADE17 were found as outliers in this analysis and in the published *elp3Δ*/*urm1Δ* analysis. The *URM1* deletion was also observed, whereas the expected *ELP3* deletion was not observed because elp3 was not quantified in this study. Rtc3, Sds24, Hxk1, Shm2, Asn1, and Ald5 were observed in this study, but in contrast to Rezgui et al. (2013), did not differ between datasets. Many of the protein changes are still directionally consistent between mutants. Filtering these proteins to those that change in the opposite

direction, representing potentially biologically interesting differences between ELP and *URM1*, reveals only 16 proteins which do not result in any significant GO categories.

It should be cautioned that this comparison between *elp3Δ*/wt and *urm1Δ*/wt datasets is not a robust analysis as it lacks adequate statistics. Furthermore, a similar number of outliers could be explained as false positives due to the FDR thresholds used during the Bayes moderated t-test. We can however conclude, in agreement with Rezgui et al. (2013), that the *URM1* and ELP pathways control an overlapping set of target proteins, and that the proteome difference between *urm1Δ* and *elp3Δ* cells is likely to be less than either mutant against wt.

The fact that there are differences between *urm1Δ* and *elp3Δ* suggests that there are ELP specific functions to investigate. The differences may related to many aspects of the complex phenotype: to mcm5 modification on mcm5S2U tRNAs, mcm5S2U independent mcm5 and ncm5 modifications, or to functions not related to tRNA modifications. Given the strong relationship between *urm1Δ* and *elp3Δ* due to co-modification of mcm5S2U, a direct analysis of *elp3Δ*/*urm1Δ* is necessary to determine if additional ELP specific codon biases or other effects exist. Given the small number of significant protein changes observed in *elp3Δ*/*urm1Δ* to date, a more in depth and statistically robust analysis is required to gather sufficient proteome coverage for a codon bias analysis, which is a future research project that will benefit from the research presented here.

Systematic Name	Standard Name	logFC *elp3Δ*/wt	logFC *urm1Δ*/wt	Calculated *elp3Δ*/*urm1Δ*
R0040C	REP2	**-1.43**	0.05	-1.47
YGR286C	BIO2	**-0.51**	0.80	-1.31
YNL251C	NRD1	**-0.94**	0.02	-0.96
YMR120C	ADE17	**-0.77**	0.15	-0.92
YKL006C-A	SFT1	**-0.17**	**0.68**	-0.84
YMR275C	BUL1	**-0.64**	0.19	-0.84
YBR092C	PHO3	**-0.61**	0.16	-0.78
YER063W	THO1	**-0.36**	0.35	-0.71
YDR116C	MRPL1	**-0.23**	0.47	-0.70
YNL160W	YGP1	**-2.33**	**-1.65**	-0.68
YIL015W	BAR1	**-0.70**	-0.07	-0.63
YDR258C	HSP78	**0.92**	0.31	0.61
YER152C	YER152C	**0.75**	0.13	0.62
YJL191W	RPS14B	**0.79**	0.16	0.64
YJL052W	TDH1	-0.73	**-1.40**	0.67
YNL036W	NCE103	**0.51**	-0.16	0.67
YOR367W	SCP1	**0.54**	-0.13	0.67
YMR305C	SCW10	**0.85**	0.13	0.72
YOL058W	ARG1	**0.36**	-0.36	0.72
YJR109C	CPA2	**1.39**	**0.59**	0.80
YOL140W	ARG8	0.23	**-0.58**	0.81
YGL117W	YGL117W	**1.11**	0.25	0.86
YKL120W	OAC1	-0.50	**-1.38**	0.88
YGL028C	SCW11	**1.02**	0.12	0.89
YLR286C	CTS1	**1.21**	0.21	1.00
YHR137W	ARO9	**2.05**	**0.92**	1.13
YPR097W	YPR097W	**0.22**	-1.09	1.31
YDR380W	ARO10	**1.92**	**0.60**	1.33
YKR042W	UTH1	**1.26**	-0.11	1.38
YOL154W	ZPS1	-0.46	**-2.04**	1.59
YJL153C	INO1	-0.44	**-3.25**	2.81
YKL166C	TPK3	0.07	**-3.52**	3.59
YIL008W	URM1	**0.46**	**-3.85**	4.31

Table 4.2 List of top 4% most changing proteins between *urm1Δ* and *elp3Δ*.

Proteins listed are from Figure 4.28 (bold dots) sorted by calculated *elp3Δ*/*urm1Δ* log2 fold change. *elp3Δ* or *urm1Δ* ratios are in bold if they were found to be significantly changing in their respective studies.

Chapter V
General Discussion

5.1 SoMBIE

The SoMBIE methodology has been successfully developed for enriching isopeptides to enhance identification by mass spectrometry. The MIRO1 ubiquitylation example demonstrated the methods effectiveness and that even relatively simple *in vitro* reactions can significantly benefit from isopeptide enrichment. This method also offers the ability to enrich NEDD8 and ISG15 isopeptides, as well as N-terminal diglycine modifications. The method did display its limitations though. Due to the reactivity towards mono-glycine, the method is not applicable to highly complex samples. The niche for this method is therefore improving isopeptide sensitivity in samples of low complexity, such as *in vitro* reactions or immunoprecipitations of ubiquitylated substrate proteins or complexes.

While this may seem limiting compared to the in-depth analysis of ubiquitin isopeptides afforded by the diglycyl-lysine reactive antibodies (Kim et al., 2011; Xu et al., 2010), these analyses also have their limitations. Of the many isopeptides identified, it is still not clear which E3 ligases are responsible for mediating the conjugation. To further understand the complex network between E2, E3, and substrate specificities, more controlled experiments are required, such as an *in vitro* screening platform.

In previous *in vitro* screens, such as the E3 Rsp5p screen against a panel of expressed yeast proteins (Kus et al., 2005), substrate ubiquitylation was detected at the protein level. The coupling of mass spectrometry to an *in vitro* screen could offer an

increased screen specificity down to lysine reactivity. In conjunction with some degree of multiplexing, or use of rapid MS techniques like MALDI-TOF/TOF, future screening platforms incorporating SoMBIE could help elucidate mechanisms of specificity in the ubiquitin system.

5.2 ICID

A novel methodology has been presented for the analysis of UBL isopeptides with long remnants. An implementation has been demonstrated with isotopically heavy SUMO2 for *in vitro* SUMOylation of a well know substrate, Ran GTPase-activating protein 1 (RanGAP1), and a potentially novel substrate, RNA guanine-7 methyltransferase (RNMT). The synthesis of isotope labelled SUMO, mass spectrometry data acquisition, and bioinformatic analysis has been demonstrated to be relatively straightforward and can be achieved without the extensive MS or bioinformatic expertise formally required for analysis of complex UBL isopeptides. Although the method was only demonstrated for SUMO, it can be adapted for application to the whole family of ubiquitin like protein modifiers and is of benefit to the wider scientific community investigating UBL protein modifications.

The ICID method can also be used for diglycine remnant UBLs (ubiquitin, NEDD8, and ISG15) to aid identification of elusive ubiquitin isopeptides. In this case, ICID and SoMBIE are not exclusive technologies. Isotope coded *in vitro* ubiquitylation could be followed by enrichment of isopeptides for optimal detection and identification.

A variation of the method was also presented that holds promise for the application of the same concepts for MS2-independent isopeptide identification to *in vivo* samples.

5.3 HyperProphet

Presented was a new workflow and software for efficient and accurate quantitative SILAC analyses in unfractionated proteomes. The biological example, *elp3Δ* yeast, demonstrated we can dig deeper into a proteome while avoiding the much overused '2-fold significance threshold'. The new application, HyperProphet, offers a relatively simple integration into TPP and requires minimal expertise to operate. The final user experience is essentially identical to a typical TPP analysis and does not require mastering of additional software environments.

A key aspect for improving functionality and robustness of future versions of HyperProphet is improvement in feature detection. This could be achieved by validating the isotopic cluster during XIC time adjustments or shifting to third party feature detection algorithms such as SuperHirn (Mueller et al., 2007) or Hardklör (Hoopmann et al., 2007). A current assumption about experimental designs is that a transfer occurs only between highly similar proteomes. As analyses become less similar, the potential for false positive peptide identification transfer will likely increase. A more discerning feature detection will permit co-analyses across samples from different experiments, such as unrelated biological conditions or a composite master-map, with reduced errors. A more stringent feature detection will also allow for transfer of peptide identifications between fractionated samples. While future proteomics technologies will likely make the unfractionated proteome the experimental design of choice, until technologies are sufficiently advanced, some degree of fractionation of highly complex mammalian proteomes will still assist analyses requiring very high proteome coverage.

Another resource for managing peptide identification events across DDA acquisition analyses was very recently published (Bateman et al., 2014). In this approach, MS1 extracted ion chromatograms were directly extracted from unidentified features based on chromatographic time alignment to additional sample analyses. Functionalities for their approach will be integrated into a future release of Skyline (MacLean et al., 2010). The recent developments by Bateman and coworkers, the "match between runs" feature in MaxQuant, and Hyperprophet are very different implementations of a similar concept. This is indicative of the proteomics community moving towards a more effective use of DDA LCMS information distributed over an experiment. The research presented here not only offers a unique implementation of this concept, but also extends the application by demonstrating that the combination of single channels and SILAC datasets can have a positive impact on the quantitative proteome coverage.

A recent development in neutron encoding for SILAC labelling utilises very high resolution MS instrumentation to measure the 6 mDa difference in mass deficits between ^{13}C and ^{15}N isotopes (Hebert et al., 2013). In this 'NeuCode' approach, a moderate resolution scan permits identification of unresolved isotope precursor features, whereas a very high resolution scan allows for quantitation of resolved stable mono-isotopes. This is an elegant solution to dealing with the proteome complexity challenges being addressed by HyperProphet. The NeuCode isotope labelling approach is an example of a new technology that will likely become mainstream as very high resolution MS instrumentation become common place. Importantly, while developments like NeuCode and HyperProphet are both dealing with proteome complexity in very different ways, they are not exclusive technologies. I foresee that combinations of technologies - isotope coding reagents, proteome 'master maps', targeted proteomics, novel sample acquisition techniques, coupled with rapidly

advancing MS hardware - will bring us one step closer to the single-shot complete proteome. The near future will bring us within a stones throw of proteome platforms akin to the high throughput DNA sequencing platforms that evolved decades before us, and with it will come more ambitious experiments to explore the complexities of biological systems.

References

Agris, P.F. (2004). Decoding the genome: a modified view. Nucleic Acids Res *32*, 223–238.

Anderson, S.L., Coli, R., Daly, I.W., Kichula, E.A., Rork, M.J., Volpi, S.A., Ekstein, J., and Rubin, B.Y. (2001). Familial dysautonomia is caused by mutations of the IKAP gene. Am. J. Hum. Genet. *68*, 753–758.

Aregger, M., and Cowling, V.H. (2013). Human cap methyltransferase (RNMT) N-terminal non-catalytic domain mediates recruitment to transcription initiation sites. Biochem J *455*, 67–73.

Atanassov, I., and Urlaub, H. (2013). Increased proteome coverage by combining PAGE and peptide isoelectric focusing: comparative study of gel-based separation approaches. Proteomics *13*, 2947–2955.

Bateman, N.W., Goulding, S.P., Shulman, N.J., Gadok, A.K., Szumlinski, K.K., Maccoss, M.J., and Wu, C.C. (2014). Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (DDA). Mol Cell Proteomics *13*, 329–338.

Bauer, F., Matsuyama, A., Candiracci, J., Dieu, M., Scheliga, J., Wolf, D.A., Yoshida, M., and Hermand, D. (2012). Translational control of cell division by Elongator. Cell Rep *1*, 424–433.

Becker, J., Barysch, S.V., Karaca, S., Dittner, C., Hsiao, H.-H., Berriel Diaz, M., Herzig, S., Urlaub, H., and Melchior, F. (2013). Detecting endogenous SUMO targets in mammalian cells and tissues. Nat. Struct. Mol. Biol. *20*, 525–531.

Begley, U., Dyavaiah, M., Patil, A., Rooney, J.P., DiRenzo, D., Young, C.M., Conklin, D.S., Zitomer, R.S., and Begley, T.J. (2007). Trm9-catalyzed tRNA modifications link translation to the DNA damage response. Mol Cell *28*, 860–870.

Bennett, E.J., Rush, J., Gygi, S.P., and Harper, J.W. (2010). Dynamics of Cullin-RING Ubiquitin Ligase Network Revealed by Systematic Quantitative Proteomics. Cell *143*, 951–965.

Blomster, H.A., Hietakangas, V., Wu, J., Kouvonen, P., Hautaniemi, S., and Sistonen, L. (2009). Novel proteomics strategy brings insight into the prevalence of SUMO-2 target sites. Mol Cell Proteomics *8*, 1382–1390.

Blomster, H.A., Imanishi, S.Y., Siimes, J., Kastu, J., Morrice, N.A., Eriksson, J.E., and Sistonen, L. (2010). In vivo identification of sumoylation sites by a signature tag and cysteine-targeted affinity purification. J Biol Chem *285*, 19324–19329.

Bodenmiller, B., Mueller, L.N., Mueller, M., Domon, B., and Aebersold, R. (2007). Reproducible isolation of distinct, overlapping segments of the phosphoproteome. Nat Meth *4*, 231–237.

Bongaerts, J., Krämer, M., Müller, U., Raeven, L., and Wubbolts, M. (2001). Metabolic engineering for microbial production of aromatic amino acids and derived compounds.

Metab. Eng. *3*, 289–300.

Brittain, S.M., Ficarro, S.B., Brock, A., and Peters, E.C. (2005). Enrichment and analysis of peptide subsets using fluorous affinity tags and mass spectrometry. Nat Biotechnol *23*, 463–468.

Bruderer, R., Tatham, M.H., Plechanovová, A., Matic, I., Garg, A.K., and Hay, R.T. (2011). Purification and identification of endogenous polySUMO conjugates. EMBO Rep *12*, 142–148.

Burande, C.F., Heuzé, M.L., Lamsoul, I., Monsarrat, B., Uttenweiler-Joseph, S., and Lutz, P.G. (2009). A label-free quantitative proteomics strategy to identify E3 ubiquitin ligase substrates targeted to proteasome degradation. Mol Cell Proteomics *8*, 1719–1727.

Butter, F., Bucerius, F., Michel, M., Cicova, Z., Mann, M., and Janzen, C.J. (2013). Comparative proteomics of two life cycle stages of stable isotope-labeled Trypanosoma brucei reveals novel components of the parasite's host adaptation machinery. Mol Cell Proteomics *12*, 172–179.

Carrano, A.C., and Bennett, E.J. (2013). Using the ubiquitin-modified proteome to monitor protein homeostasis function. Mol Cell Proteomics *12*, 3521–3531.

Chan, L., Cross, H.F., She, J.K., Cavalli, G., Martins, H.F.P., and Neylon, C. (2007). Covalent attachment of proteins to solid supports and surfaces via Sortase-mediated ligation. PLoS ONE *2*, e1164.

Chen, C., Huang, B., Anderson, J.T., and BYSTRÖM, A.S. (2011a). Unexpected accumulation of ncm(5)U and ncm(5)S(2) (U) in a trm9 mutant suggests an additional step in the synthesis of mcm(5)U and mcm(5)S(2)U. PLoS ONE *6*, e20783.

Chen, C., Tuck, S., and BYSTRÖM, A.S. (2009). Defects in tRNA modification associated with neurological and developmental dysfunctions in Caenorhabditis elegans elongator mutants. PLoS Genet *5*, e1000561.

Chen, I., Dorr, B.M., and Liu, D.R. (2011b). A general strategy for the evolution of bond-forming enzymes using yeast display. Proc Natl Acad Sci USA *108*, 11399–11404.

Chen, P.-C., Na, C.H., and Peng, J. (2012). Quantitative proteomics to decipher ubiquitin signaling. Amino Acids *43*, 1049–1060.

Chicooree, N., Connolly, Y., Tan, C.-T., Malliri, A., Li, Y., Smith, D.L., and Griffiths, J.R. (2013a). Enhanced detection of ubiquitin isopeptides using reductive methylation. J Am Soc Mass Spectrom *24*, 421–430.

Chicooree, N., Griffiths, J.R., Connolly, Y., and Smith, D.L. (2013b). Chemically facilitating the generation of diagnostic ions from SUMO(2/3) remnant isopeptides. Rapid Commun Mass Spectrom *27*, 2108–2114.

Chicooree, N., Griffiths, J.R., Connolly, Y., Tan, C.-T., Malliri, A., Eyers, C.E., and Smith, D.L. (2013c). A novel approach to the analysis of SUMOylation with the independent use of trypsin and elastase digestion followed by database searching utilising consecutive residue addition to lysine. Rapid Commun Mass Spectrom *27*,

127–134.

Chiou, H.-Y.C., Liu, S.-Y., Lin, C.-H., and Lee, E.H. (2014). Hes-1 SUMOylation by protein inhibitor of activated STAT1 enhances the suppressing effect of Hes-1 on GADD45α expression to increase cell survival. J. Biomed. Sci. *21*, 53.

Clauser, K.R., Baker, P., and Burlingame, A.L. (1999). Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal Chem *71*, 2871–2882.

Cohen, L., Henzel, W.J., and Baeuerle, P.A. (1998). IKAP is a scaffold protein of the IkappaB kinase complex. Nature *395*, 292–296.

Colaert, N., Van Huele, C., Degroeve, S., Staes, A., Vandekerckhove, J., Gevaert, K., and Martens, L. (2011). Combining quantitative proteomics data processing workflows for greater sensitivity. Nat Meth *8*, 481–483.

Cooper, H.J., Tatham, M.H., Jaffray, E., Heath, J.K., Lam, T.T., Marshall, A.G., and Hay, R.T. (2005). Fourier transform ion cyclotron resonance mass spectrometry for the analysis of small ubiquitin-like modifier (SUMO) modification: identification of lysines in RanBP2 and SUMO targeted for modification during the E3 autoSUMOylation reaction. Anal Chem *77*, 6310–6319.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. Science *327*, 425–431.

Cowling, V.H. (2010). Regulation of mRNA cap methylation. Biochem J *425*, 295–302.

Cox, J.J., and Mann, M.M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol *26*, 1367–1372.

Craig, R., and Beavis, R.C. (2004). TANDEM: matching proteins with tandem mass spectra. Bioinformatics *20*, 1466–1467.

Creppe, C., and Buschbeck, M. (2011). Elongator: an ancestral complex driving transcription and migration through protein acetylation. J. Biomed. Biotechnol. *2011*, 924898.

Creppe, C., Malinouskaya, L., Volvert, M.-L., Gillard, M., Close, P., Malaise, O., Laguesse, S., Cornez, I., Rahmouni, S., Ormenese, S., et al. (2009). Elongator controls the migration and differentiation of cortical neurons through acetylation of alpha-tubulin. Cell *136*, 551–564.

Crick, F.H. (1966). Codon--anticodon pairing: the wobble hypothesis. J Mol Biol *19*, 548–555.

Cristobal, A., Hennrich, M.L., Giansanti, P., Goerdayal, S.S., Heck, A.J.R., and Mohammed, S. (2012). In-house construction of a UHPLC system enabling the identification of over 4000 protein groups in a single analysis. Analyst *137*, 3541–3548.

Danielsen, J.M.R., Sylvestersen, K.B., Bekker-Jensen, S., Szklarczyk, D., Poulsen, J.W., Horn, H., Jensen, L.J., Mailand, N., and Nielsen, M.L. (2011). Mass spectrometric

analysis of lysine ubiquitylation reveals promiscuity at site level. Mol Cell Proteomics *10*, M110.003590.

David, Y., Ternette, N., Edelmann, M.J., Ziv, T., Gayer, B., Sertchook, R., Dadon, Y., Kessler, B.M., and Navon, A. (2011). E3 ligases determine ubiquitination site and conjugate type by enforcing specificity on E2 enzymes. J Biol Chem *286*, 44104–44115.

de Godoy, L.M.F., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Fröhlich, F., Walther, T.C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature *455*, 1251–1254.

Delmotte, N., Lasaosa, M., Tholey, A., Heinzle, E., and Huber, C.G. (2007). Two-dimensional reversed-phase x ion-pair reversed-phase HPLC: an alternative approach to high-resolution peptide separation for shotgun proteome analysis. J Proteome Res *6*, 4363–4373.

Denis, N.J., Vasilescu, J., Lambert, J.-P., Smith, J.C., and Figeys, D. (2007). Tryptic digestion of ubiquitin standards reveals an improved strategy for identifying ubiquitinated proteins by mass spectrometry. Proteomics *7*, 868–874.

Denison, C., Rudner, A.D., Gerber, S.A., Bakalarski, C.E., Moazed, D., and Gygi, S.P. (2005). A proteomic strategy for gaining insights into protein sumoylation in yeast. Mol Cell Proteomics *4*, 246–254.

Desterro, J.M., Rodriguez, M.S., and Hay, R.T. (1998). SUMO-1 modification of IkappaBalpha inhibits NF-kappaB activation. Mol Cell *2*, 233–239.

Dramsi, S., Trieu-Cuot, P., and Bierne, H. (2005). Sorting sortases: a nomenclature proposal for the various sortases of Gram-positive bacteria. Research in Microbiology *156*, 289–297.

Dumont, Q., Donaldson, D.L., and Griffith, W.P. (2011). Screening method for isopeptides from small ubiquitin-related modifier-conjugated proteins by ion mobility mass spectrometry. Anal Chem *83*, 9638–9642.

Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Jam *5*, 976–989.

Eng, J.K., Jahan, T.A., and Hoopmann, M.R. (2013). Comet: an open-source MS/MS sequence database search tool. Proteomics *13*, 22–24.

Esberg, A., Huang, B., JOHANSSON, M.J.O., and BYSTRÖM, A.S. (2006). Elevated levels of two tRNA species bypass the requirement for elongator complex in transcription and exocytosis. Mol Cell *24*, 139–148.

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., and Whitehouse, C.M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. Science *246*, 64–71.

Flotho, A., Werner, A., Winter, T., Frank, A.S., Ehret, H., and Melchior, F. (2012). Recombinant reconstitution of sumoylation reactions in vitro. Methods Mol Biol *832*, 93–110.

Fuhs, S.R., and Insel, P.A. (2011). Caveolin-3 undergoes SUMOylation by the SUMO E3 ligase PIASy: sumoylation affects G-protein-coupled receptor desensitization. J Biol Chem *286*, 14830–14841.

Furukawa, K., Mizushima, N., Noda, T., and Ohsumi, Y. (2000). A protein conjugation system in yeast with homology to biosynthetic enzyme reaction of prokaryotes. J Biol Chem *275*, 7462–7465.

Galisson, F., Mahrouche, L., Courcelles, M., Bonneil, E., Meloche, S., Chelbi-Alix, M.K., and Thibault, P. (2011). A novel proteomics approach to identify SUMOylated proteins and their modification sites in human cells. Molecular & Cellular Proteomics *10*, M110.004796.

Gatto, L., and Christoforou, A. (2014). Using R and Bioconductor for proteomics data analysis. Biochim Biophys Acta *1844*, 42–51.

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631–636.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics *11*, M111.014050.

Goehring, A.S., Rivers, D.M., and Sprague, G.F. (2003). Attachment of the ubiquitin-related protein Urm1p to the antioxidant protein Ahp1p. Eukaryotic Cell *2*, 930–936.

Goldstein, G., Scheid, M., Hammerling, U., Schlesinger, D.H., Niall, H.D., and Boyse, E.A. (1975). Isolation of a polypeptide that has lymphocyte-differentiating properties and is probably represented universally in living cells. Proc Natl Acad Sci USA *72*, 11–15.

Golebiowski, F., Matic, I., Tatham, M.H., Cole, C., Yin, Y., Nakamura, A., Cox, J., Barton, G.J., Mann, M., and Hay, R.T. (2009). System-wide changes to SUMO modifications in response to heat shock. Science Signaling *2*, ra24.

Golebiowski, F., Tatham, M.H., Nakamura, A., and Hay, R.T. (2010). High-stringency tandem affinity purification of proteins conjugated to ubiquitin-like moieties. Nat Protoc *5*, 873–882.

Gonatopoulos-Pournatzis, T., Dunn, S., Bounds, R., and Cowling, V.H. (2011). RAM/Fam103a1 is required for mRNA cap methylation. Mol Cell *44*, 585–596.

Gotelli, N.J., Ellison, A.M., and Ballif, B.A. (2012). Environmental proteomics, biodiversity statistics and food-web structure. Trends Ecol. Evol. (Amst.) *27*, 436–442.

Gresch, O., and Altrogge, L. (2012). Transfection of difficult-to-transfect primary mammalian cells. Methods Mol Biol *801*, 65–74.

Hahne, H., Pachl, F., Ruprecht, B., Maier, S.K., and Klaeger, S. (2013). DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. Nature.

Han, D.K., Eng, J., Zhou, H., and Aebersold, R. (2001). Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass

spectrometry. Nat Biotechnol *19*, 946–951.

Hebert, A.S., Merrill, A.E., Bailey, D.J., Still, A.J., Westphall, M.S., Strieter, E.R., Pagliarini, D.J., and Coon, J.J. (2013). Neutron-encoded mass signatures for multiplexed proteome quantification. Nat Meth *10*, 332–334.

Hebert, A.S., Richards, A.L., Bailey, D.J., Ulbrich, A., Coughlin, E.E., Westphall, M.S., and Coon, J.J. (2014). The one hour yeast proteome. Mol Cell Proteomics *13*, 339–347.

Hitchcock, A.L., Auld, K., Gygi, S.P., and Silver, P.A. (2003). A subset of membrane-associated proteins is ubiquitinated in response to mutations in the endoplasmic reticulum degradation machinery. Proc Natl Acad Sci USA *100*, 12735–12740.

Hjerpe, R., Aillet, F., Lopitz-Otsoa, F., Lang, V., England, P., and Rodriguez, M.S. (2009). Efficient protection and isolation of ubiquitylated proteins using tandem ubiquitin-binding entities. EMBO Rep *10*, 1250–1258.

Hjerpe, R., Thomas, Y., Chen, J., Zemla, A., Curran, S., Shpiro, N., Dick, L.R., and Kurz, T. (2012). Changes in the ratio of free NEDD8 to ubiquitin triggers NEDDylation by ubiquitin enzymes. Biochem J *441*, 927–936.

Hoopmann, M.R., Finney, G.L., and Maccoss, M.J. (2007). High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. Anal Chem *79*, 5620–5632.

Hoopmann, M.R., Merrihew, G.E., Haller, von, P.D., and Maccoss, M.J. (2009). Post analysis data acquisition for the iterative MS/MS sampling of proteomics mixtures. J Proteome Res *8*, 1870–1875.

Hörth, P., Miller, C.A., Preckel, T., and Wenz, C. (2006). Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis. Mol Cell Proteomics *5*, 1968–1974.

Hsiao, H.-H., Meulmeester, E., Frank, B.T.C., Melchior, F., and Urlaub, H. (2009). "ChopNSpice," a mass spectrometric approach that allows identification of endogenous small ubiquitin-like modifier-conjugated peptides. Mol Cell Proteomics *8*, 2664–2675.

Huang, B., JOHANSSON, M.J.O., and BYSTRÖM, A.S. (2005). An early step in wobble uridine tRNA modification requires the Elongator complex. Rna *11*, 424–436.

Huang, B., Lu, J., and BYSTRÖM, A.S. (2008). A genome-wide screen identifies genes required for formation of the wobble nucleoside 5-methoxycarbonylmethyl-2-thiouridine in Saccharomyces cerevisiae. Rna *14*, 2183–2194.

Huang, X., Aulabaugh, A., Ding, W., Kapoor, B., Alksne, L., Tabei, K., and Ellestad, G. (2003). Kinetic mechanism of Staphylococcus aureus sortase SrtA. Biochemistry *42*, 11307–11315.

Iben, J.R., and Maraia, R.J. (2012). tRNAomics: tRNA gene copy number variation and codon use provide bioinformatic evidence of a new anticodon:codon wobble pair in a eukaryote. Rna *18*, 1358–1372.

Ikeda, F., and Dikic, I. (2008). Atypical ubiquitin chains: new molecular signals. "Protein Modifications: Beyond the Usual Suspects" review series. EMBO Rep *9*, 536–

542.

Issaq, H.J., Conrads, T.P., Janini, G.M., and Veenstra, T.D. (2002). Methods for fractionation, separation and profiling of proteins and peptides. Electrophoresis *23*, 3048–3061.

Jeram, S.M., Srikumar, T., Pedrioli, P.G.A., and Raught, B. (2009). Using mass spectrometry to identify ubiquitin and ubiquitin-like protein conjugation sites. Proteomics *9*, 922–934.

Jeram, S.M., Srikumar, T., Zhang, X.-D., Anne Eisenhauer, H., Rogers, R., Pedrioli, P.G.A., Matunis, M., and Raught, B. (2010). An improved SUMmOn-based methodology for the identification of ubiquitin and ubiquitin-like protein conjugation sites identifies novel ubiquitin-like protein chain linkages. Proteomics *10*, 254–265.

JOHANSSON, M.J.O., Esberg, A., Huang, B., Björk, G.R., and BYSTRÖM, A.S. (2008). Eukaryotic wobble uridine modifications promote a functionally redundant decoding system. Mol Cell Biol *28*, 3301–3312.

Johnson, E.S., and Gupta, A.A. (2001). An E3-like factor that promotes SUMO conjugation to the yeast septins. Cell *106*, 735–744.

Kalli, A., Smith, G.T., Sweredoski, M.J., and Hess, S. (2013). Evaluation and Optimization of Mass Spectrometric Settings during Data-Dependent Acquisition Mode: Focus on LTQ-Orbitrap Mass Analyzers. J Proteome Res.

Karas, M., and Hillenkamp, F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem *60*, 2299–2301.

Kazlauskaite, A., van kelly, Johnson, C., Baillie, C., Hastie, C.J., Peggie, M., Macartney, T., Woodroof, H.I., Alessi, D.R., Pedrioli, P.G.A., et al. (2014). Phosphorylation of Parkin at Serine65 is essential for activation: elaboration of a Miro1 substrate-based assay of Parkin E3 ligase activity. Open Biol *4*, 130213–130213.

Keith, G. (1984). The primary structures of two arginine tRNAs (anticodons C-C-U and mcm5a2U-C-psi) and of glutamine tRNA (anticodon C-U-G) from bovine liver. Nucleic Acids Res *12*, 2543–2547.

Keller, A., Eng, J., Zhang, N., Li, X.-J., and Aebersold, R. (2005). A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol. Syst. Biol. *1*, 2005.0017.

Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem *74*, 5383–5392.

Kelstrup, C.D., Young, C., Lavallee, R., Nielsen, M.L., and Olsen, J.V. (2012). Optimized fast and sensitive acquisition methods for shotgun proteomics on a quadrupole orbitrap mass spectrometer. J Proteome Res *11*, 3487–3497.

Kim, W., Bennett, E.J., Huttlin, E.L., Guo, A., Li, J., Possemato, A., Sowa, M.E., Rad, R., Rush, J., Comb, M.J., et al. (2011). Systematic and quantitative assessment of the ubiquitin-modified proteome. Mol Cell *44*, 325–340.

Kirkpatrick, D.S., Denison, C., and Gygi, S.P. (2005). Weighing in on ubiquitin: the expanding role of mass-spectrometry-based proteomics. Nat Cell Biol *7*, 750–757.

Klug, H., Xaver, M., Chaugule, V.K., Koidl, S., Mittler, G., Klein, F., and Pichler, A. (2013). Ubc9 sumoylation controls SUMO chain formation and meiotic synapsis in Saccharomyces cerevisiae. Mol Cell *50*, 625–636.

Knuesel, M., Cheung, H.T., Hamady, M., Barthel, K.K.B., and Liu, X. (2005). A method of mapping protein sumoylation sites by mass spectrometry using a modified small ubiquitin-like modifier 1 (SUMO-1) and a computational program. Mol Cell Proteomics *4*, 1626–1636.

Kohlbacher, O., Reinert, K., Gröpl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., and Sturm, M. (2007). TOPP--the OpenMS proteomics pipeline. Bioinformatics *23*, e191–e197.

Kondapalli, C., Kazlauskaite, A., Zhang, N., Woodroof, H.I., Campbell, D.G., Gourlay, R., Burchell, L., Walden, H., Macartney, T.J., Deak, M., et al. (2012). PINK1 is activated by mitochondrial membrane potential depolarization and stimulates Parkin E3 ligase activity by phosphorylating Serine 65. Open Biol *2*, 120080.

Kraft, C., Peter, M., and Hofmann, K. (2010). Selective autophagy: ubiquitin-mediated recognition and beyond. Nat Cell Biol *12*, 836–841.

Kriegenburg, F., Ellgaard, L., and Hartmann-Petersen, R. (2012). Molecular chaperones in targeting misfolded proteins for ubiquitin-dependent degradation. Febs J. *279*, 532–542.

Krogan, N.J., and Greenblatt, J.F. (2001). Characterization of a six-subunit holo-elongator complex required for the regulated expression of a group of genes in Saccharomyces cerevisiae. Mol Cell Biol *21*, 8203–8212.

Krokhin, O.V., and Spicer, V. (2009). Peptide retention standards and hydrophobicity indexes in reversed-phase high-performance liquid chromatography of peptides. Anal Chem *81*, 9522–9530.

Kruger, R.G., Otvos, B., Frankel, B.A., Bentley, M., Dostal, P., and McCafferty, D.G. (2004). Analysis of the substrate specificity of the Staphylococcus aureus sortase transpeptidase SrtA. Biochemistry *43*, 1541–1551.

Kumar, S., Yoshida, Y., and Noda, M. (1993). Cloning of a cDNA Which Encodes a Novel Ubiquitin-like Protein. Biochem Biophys Res Commun *195*, 393–399.

Kus, B., Gajadhar, A., Stanger, K., Cho, R., Sun, W., Rouleau, N., Lee, T., Chan, D., Wolting, C., Edwards, A., et al. (2005). A high throughput screen to identify substrates for the ubiquitin ligase Rsp5. J Biol Chem *280*, 29470–29478.

Lamoliatte, F., Bonneil, E., Durette, C., Caron-Lizotte, O., wildemann, D., Zerweck, J., Wenshuk, H., and Thibault, P. (2013). Targeted identification of SUMOylation sites in human proteins using affinity enrichment and paralog-specific reporter ions. Mol Cell Proteomics *12*, 2536–2550.

Lauwers, E., Jacob, C., and André, B. (2009). K63-linked ubiquitin chains as a specific signal for protein sorting into the multivesicular body pathway. J. Cell Biol. *185*, 493–

502.

Leidel, S., Pedrioli, P.G.A., Bucher, T., Brost, R., Costanzo, M., Schmidt, A., Aebersold, R., Boone, C., Hofmann, K., and Peter, M. (2009). Ubiquitin-related modifier Urm1 acts as a sulphur carrier in thiolation of eukaryotic transfer RNA. Nature *458*, 228–232.

Leng, L., Xu, C., Wei, C., Zhang, J., Liu, B., Ma, J., Li, N., Qin, W., Zhang, W., Zhang, C., et al. (2013). A Proteomics Strategy for the Identification of FAT10-Modified Sites by Mass Spectrometry. J Proteome Res.

Levary, D.A., Parthasarathy, R., Boder, E.T., and Ackerman, M.E. (2011). Protein-protein fusion catalyzed by sortase A. PLoS ONE *6*, e18342.

Levin, Y. (2011). The role of statistical power analysis in quantitative proteomics. Proteomics *11*, 2565–2567.

Li, W., Bengtson, M.H., Ulbrich, A., Matsuda, A., Reddy, V.A., Orth, A., Chanda, S.K., Batalov, S., and Joazeiro, C.A.P. (2008). Genome-wide and functional annotation of human E3 ubiquitin ligases identifies MULAN, a mitochondrial E3 that regulates the organelle's dynamics and signaling. PLoS ONE *3*, e1487.

Li, Z., and Li, L. (2014). Chemical-vapor-assisted electrospray ionization for increasing analyte signals in electrospray ionization mass spectrometry. Anal Chem *86*, 331–335.

Liakopoulos, D., Doenges, G., Matuschewski, K., and Jentsch, S. (1998). A novel protein modification pathway related to the ubiquitin system. Embo J *17*, 2208–2214.

Lin, F.-J., Shen, L., Jang, C.-W., Falnes, P.Ø., and Zhang, Y. (2013). Ikbkap/Elp1 deficiency causes male infertility by disrupting meiotic progression. PLoS Genet *9*, e1003516.

Lin, M.T., Sperling, L.J., Frericks Schmidt, H.L., Tang, M., Samoilova, R.I., Kumasaka, T., Iwasaki, T., Dikanov, S.A., Rienstra, C.M., and Gennis, R.B. (2011). A rapid and robust method for selective isotope labeling of proteins. Methods *55*, 370–378.

Lo, A., Weiner, J.H., and Li, L. (2013). Analytical performance of reciprocal isotope labeling of proteome digests for quantitative proteomics and its application for comparative studies of aerobic and anaerobic Escherichia coli proteomes. Anal Chim Acta *795*, 25–35.

MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., Kern, R., Tabb, D.L., Liebler, D.C., and Maccoss, M.J. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics *26*, 966–968.

Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem *72*, 1156–1162.

Marko-Varga, G., Ogiwara, A., Nishimura, T., Kawamura, T., Fujii, K., Kawakami, T., Kyono, Y., Tu, H.-K., Anyoji, H., Kanazawa, M., et al. (2007). Personalized medicine and proteomics: lessons from non-small cell lung cancer. J Proteome Res *6*, 2925–2935.

Marraffini, L.A., Dedent, A.C., and Schneewind, O. (2006). Sortases and the art of

anchoring proteins to the envelopes of gram-positive bacteria. Microbiol Mol Biol Rev *70*, 192–221.

Matic, I., Jaffray, E.G., Oxenham, S.K., Groves, M.J., Barratt, C.L.R., Tauro, S., Stanley-Wall, N.R., and Hay, R.T. (2011). Absolute SILAC-compatible expression strain allows Sumo-2 copy number determination in clinical samples. J Proteome Res *10*, 4869–4875.

Matic, I., Schimmel, J., Hendriks, I.A., van Santen, M.A., van de Rijke, F., van Dam, H., Gnad, F., Mann, M., and Vertegaal, A.C.O. (2010). Site-specific identification of SUMO-2 targets in cells reveals an inverted SUMOylation motif and a hydrophobic cluster SUMOylation motif. Mol Cell *39*, 641–652.

Matic, I., van Hagen, M., Schimmel, J., Macek, B., Ogg, S.C., Tatham, M.H., Hay, R.T., Lamond, A.I., Mann, M., and Vertegaal, A.C.O. (2008). In vivo identification of human small ubiquitin-like modifier polymerization sites by high accuracy mass spectrometry and an in vitro to in vivo strategy. Mol Cell Proteomics *7*, 132–144.

Matsumoto, M., Hatakeyama, S., Oyamada, K., Oda, Y., Nishimura, T., and Nakayama, K.I. (2005). Large-scale analysis of the human ubiquitin-related proteome. Proteomics *5*, 4145–4151.

Mattiroli, F., and Sixma, T.K. (2014). Lysine-targeting specificity in ubiquitin and ubiquitin-like modification pathways. Nat. Struct. Mol. Biol. *21*, 308–316.

Matunis, M.J., Wu, J., and Blobel, G. (1998). SUMO-1 modification and its role in targeting the Ran GTPase-activating protein, RanGAP1, to the nuclear pore complex. J. Cell Biol. *140*, 499–509.

May, D., Law, W., Fitzgibbon, M., Fang, Q., and McIntosh, M. (2009). Software platform for rapidly creating computational tools for mass spectrometry-based proteomics. J Proteome Res *8*, 3212–3217.

Mayor, T., Lipford, J.R., Graumann, J., Smith, G.T., and Deshaies, R.J. (2005). Analysis of polyubiquitin conjugates reveals that the Rpn10 substrate receptor contributes to the turnover of multiple proteasome targets. Mol Cell Proteomics *4*, 741–751.

Mazmanian, S.K., Liu, G., Ton-That, H., and Schneewind, O. (1999). Staphylococcus aureus sortase, an enzyme that anchors surface proteins to the cell wall. Science *285*, 760–763.

Mehlhorn, J., Steinocher, H., Beck, S., Kennis, J.T.M., Hegemann, P., and Mathes, T. (2013). A set of engineered Escherichia coli expression strains for selective isotope and reactivity labeling of amino acid side chains and flavin cofactors. PLoS ONE *8*, e79006.

Merbl, Y., Refour, P., Patel, H., Springer, M., and Kirschner, M.W. (2013). Profiling of ubiquitin-like modifications reveals features of mitotic control. Cell *152*, 1160–1172.

Merlet, J., Burger, J., Gomes, J.-E., and Pintard, L. (2009). Regulation of cullin-RING E3 ubiquitin-ligases by neddylation and dimerization. Cell. Mol. Life Sci. *66*, 1924–1938.

Metzger, M.B., Hristova, V.A., and Weissman, A.M. (2012). HECT and RING finger

families of E3 ubiquitin ligases at a glance. J. Cell. Sci. *125*, 531–537.

Meyer, H.-J., and Rape, M. (2014). Enhanced Protein Degradation by Branched Ubiquitin Chains. Cell *157*, 910–921.

Meyer, J.G., and A Komives, E. (2012). Charge state coalescence during electrospray ionization improves peptide identification by tandem mass spectrometry. J Am Soc Mass Spectrom *23*, 1390–1399.

Michalski, A., Cox, J., and Mann, M. (2011a). More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J Proteome Res *10*, 1785–1793.

Michalski, A., Damoc, E., Hauschild, J.-P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011b). Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol Cell Proteomics *10*, M111.011015.

Mimura, S., Yamaguchi, T., Ishii, S., Noro, E., Katsura, T., Obuse, C., and Kamura, T. (2010). Cul8/Rtt101 forms a variety of protein complexes that regulate DNA damage response and transcriptional silencing. Journal of Biological Chemistry *285*, 9858–9867.

Moradian, A., Kalli, A., Sweredoski, M.J., and Hess, S. (2014). The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications. Proteomics *14*, 489–497.

Mueller, L.N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M.-Y., Vitek, O., Aebersold, R., and Müller, M. (2007). SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. Proteomics *7*, 3470–3480.

Münchbach, M., Quadroni, M., Miotto, G., and James, P. (2000). Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. Anal Chem *72*, 4047–4057.

Nagaraj, N., Kulak, N.A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012). System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. Molecular & Cellular Proteomics *11*, M111.013722.

Navarre, W.W., and Schneewind, O. (1999). Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. Microbiol Mol Biol Rev *63*, 174–229.

Neerathilingam, M., and Markley, J.L. (2010). Auto-induction medium containing glyphosate for high-level incorporation of unusual aromatic amino acids into proteins. BioTechniques *49*, 659–661.

Nelms, J., Edwards, R.M., Warwick, J., and Fotheringham, I. (1992). Novel mutations in the pheA gene of Escherichia coli K-12 which result in highly feedback inhibition-resistant variants of chorismate mutase/prephenate dehydratase. Appl. Environ. Microbiol. *58*, 2592–2598.

Nielsen, M.L., Vermeulen, M., Bonaldi, T., Cox, J., Moroder, L., and Mann, M. (2008).

Iodoacetamide-induced artifact mimics ubiquitination in mass spectrometry. Nat Meth *5*, 459–460.

O'Grady, C., Rempel, B.L., Sokaribo, A., Nokhrin, S., and Dmitriev, O.Y. (2012). One-step amino acid selective isotope labeling of proteins in prototrophic Escherichia coli strains. Anal Biochem *426*, 126–128.

Oppermann, F.S., Klammer, M., Bobe, C., Cox, J., Schaab, C., Tebbe, A., and Daub, H. (2013). Comparison of SILAC and mTRAQ quantification for phosphoproteomics on a quadrupole orbitrap mass spectrometer. J Proteome Res *12*, 4089–4100.

Osula, O., Swatkoski, S., and Cotter, R.J. (2012). Identification of protein SUMOylation sites by mass spectrometry using combined microwave-assisted aspartic acid cleavage and tryptic digestion. J Mass Spectrom *47*, 644–654.

Otero, G., Fellows, J., Li, Y., de Bizemont, T., Dirac, A.M., Gustafsson, C.M., Erdjument-Bromage, H., Tempst, P., and Svejstrup, J.Q. (1999). Elongator, a multisubunit component of a novel RNA polymerase II holoenzyme for transcriptional elongation. Mol Cell *3*, 109–118.

Pan, S., Zhang, H., Rush, J., Eng, J., Zhang, N., Patterson, D., Comb, M.J., and Aebersold, R. (2005). High throughput proteome screening for biomarker detection. Mol Cell Proteomics *4*, 182–190.

Pappin, D.J., Hojrup, P., and Bleasby, A.J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. Curr. Biol. *3*, 327–332.

Patel, A.H., Nowlan, P., Weavers, E.D., and Foster, T. (1987). Virulence of protein A-deficient and alpha-toxin-deficient mutants of Staphylococcus aureus isolated by allele replacement. Infect. Immun. *55*, 3103–3110.

Pedrioli, P.G.A. (2010). Trans-proteomic pipeline: a pipeline for proteomic analysis. Methods Mol Biol *604*, 213–238.

Pedrioli, P.G.A., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R.H., Apweiler, R., et al. (2004). A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol *22*, 1459–1466.

Pedrioli, P.G.A., Leidel, S., and Hofmann, K. (2008). Urm1 at the crossroad of modifications. “Protein Modifications: Beyond the Usual Suspects” Review Series. EMBO Rep *9*, 1196–1202.

Pedrioli, P.G.A., Raught, B., Zhang, X.-D., Rogers, R., Aitchison, J., Matunis, M., and Aebersold, R. (2006). Automated identification of SUMOylation sites using mass spectrometry and SUMmOn pattern recognition software. Nat Meth *3*, 533–539.

Peng, J., Schwartz, D., Elias, J.E., Thoreen, C.C., Cheng, D., Marsischky, G., Roelofs, J., Finley, D., and Gygi, S.P. (2003). A proteomics approach to understanding protein ubiquitination. Nat Biotechnol *21*, 921–926.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis *20*, 3551–3567.

Petroski, M.D. (2008). The ubiquitin system, disease, and drug discovery. BMC Biochem *9 Suppl 1*, S7.

Petroski, M.D., and Deshaies, R.J. (2005). Function and regulation of cullin-RING ubiquitin ligases. Nat Rev Mol Cell Biol *6*, 9–20.

Pirmoradian, M., Budamgunta, H., Chingin, K., Zhang, B., Astorga-Wells, J., and Zubarev, R.A. (2013). Rapid and deep human proteome analysis by single-dimension shotgun proteomics. Mol Cell Proteomics *12*, 3330–3338.

Popp, M.W., Dougan, S.K., Chuang, T.-Y., Spooner, E., and Ploegh, H.L. (2011). Sortase-catalyzed transformations that improve the properties of cytokines. Proc Natl Acad Sci USA *108*, 3169–3174.

Poulsen, J.W., Madsen, C.T., Young, C., Kelstrup, C.D., Grell, H.C., Henriksen, P., Juhl-Jensen, L., and Nielsen, M.L. (2012). Comprehensive profiling of proteome changes upon sequential deletion of deubiquitylating enzymes. Journal of Proteomics *75*, 3886–3897.

Pritz, S., Wolf, Y., Kraetke, O., Klose, J., Bienert, M., and Beyermann, M. (2007). Synthesis of biologically active peptide nucleic acid-peptide conjugates by sortase-mediated ligation. J. Org. Chem. *72*, 3909–3912.

Proft, T. (2010). Sortase-mediated protein ligation: an emerging biotechnology tool for protein modification and immobilisation. Biotechnol. Lett. *32*, 1–10.

Prudden, J., Pebernard, S., Raffa, G., Slavin, D.A., Perry, J.J.P., Tainer, J.A., McGowan, C.H., and Boddy, M.N. (2007). SUMO-targeted ubiquitin ligases in genome stability. Embo J *26*, 4089–4101.

Race, P.R., Bentley, M.L., Melvin, J.A., Crow, A., Hughes, R.K., Smith, W.D., Sessions, R.B., Kehoe, M.A., Mccafferty, D.G., and Banfield, M.J. (2009). Crystal structure of Streptococcus pyogenes sortase A: implications for sortase mechanism. J Biol Chem *284*, 6924–6933.

Rahl, P.B., Chen, C.Z., and Collins, R.N. (2005). Elp1p, the yeast homolog of the FD disease syndrome protein, negatively regulates exocytosis independently of transcriptional elongation. Mol Cell *17*, 841–853.

Reddy Chichili, V.P., Kumar, V., and Sivaraman, J. (2013). Linkers in the structural biology of protein-protein interactions. Protein Sci *22*, 153–167.

Reiter, L., Claassen, M., Schrimpf, S.P., Jovanovic, M., Schmidt, A., Buhmann, J.M., Hengartner, M.O., and Aebersold, R. (2009). Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Molecular & Cellular Proteomics *8*, 2405–2417.

Rezgui, V.A.N., Tyagi, K., Ranjan, N., Konevega, A.L., Mittelstaet, J., Rodnina, M.V., Peter, M., and Pedrioli, P.G.A. (2013). tRNA tKUUU, tQUUG, and tEUUC wobble position modifications fine-tune protein translation by promoting ribosome A-site binding. Proc Natl Acad Sci USA *110*, 12289–12294.

Ritorto, M.S., Cook, K., Tyagi, K., Pedrioli, P.G.A., and Trost, M. (2013). Hydrophilic strong anion exchange (hSAX) chromatography for highly orthogonal peptide

separation of complex proteomes. J Proteome Res *12*, 2449–2457.

Robinson, M.D., Grigull, J., Mohammad, N., and Hughes, T.R. (2002). FunSpec: a web-based cluster interpreter for yeast. BMC Bioinformatics *3*, 35.

Rodriguez, J., Gupta, N., Smith, R.D., and Pevzner, P.A. (2008). Does trypsin cut before proline? J Proteome Res *7*, 300–305.

Rodriguez, M.S., Dargemont, C., and Hay, R.T. (2001). SUMO-1 conjugation in vivo requires both a consensus modification motif and nuclear targeting. J Biol Chem *276*, 12654–12659.

Sampson, D.A., Wang, M., and Matunis, M.J. (2001). The small ubiquitin-like modifier-1 (SUMO-1) consensus sequence mediates Ubc9 binding and is essential for SUMO-1 modification. J Biol Chem *276*, 21664–21669.

Savitski, M.M., Fischer, F., Mathieson, T., Sweetman, G., Lang, M., and Bantscheff, M. (2010). Targeted data acquisition for improved reproducibility and robustness of proteomic mass spectrometry assays. J Am Soc Mass Spectrom *21*, 1668–1679.

Schmidt, A., Gehlenborg, N., Bodenmiller, B., Mueller, L.N., Campbell, D., Mueller, M., Aebersold, R., and Domon, B. (2008). An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. Mol Cell Proteomics *7*, 2138–2150.

Shi, Y., Chan, D.W., Jung, S.Y., Malovannaya, A., Wang, Y., and Qin, J. (2011a). A data set of human endogenous protein ubiquitination sites. Mol Cell Proteomics *10*, M110.002089.

Shi, Y., Xu, P., and Qin, J. (2011b). Ubiquitinated proteome: ready for global? Mol Cell Proteomics *10*, R110.006882.

Spence, J., Sadis, S., Haas, A.L., and Finley, D. (1995). A ubiquitin mutant with specific defects in DNA repair and multiubiquitination. Mol Cell Biol *15*, 1265–1273.

Sternsdorf, T., Jensen, K., Reich, B., and Will, H. (1999). The nuclear dot protein sp100, characterization of domains necessary for dimerization, subcellular localization, and modification by small ubiquitin-like modifiers. J Biol Chem *274*, 12555–12566.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. BMC Bioinformatics *9*, 307.

Strohalm, M., Kavan, D., Novak, P., Volný, M., and Havlícek, V. (2010). mMass 3: a cross-platform software environment for precise analysis of mass spectrometric data. Anal Chem *82*, 4648–4651.

Sun, H., Leverson, J.D., and Hunter, T. (2007). Conserved function of RNF4 family proteins in eukaryotes: targeting a ubiquitin ligase to SUMOylated proteins. Embo J *26*, 4102–4112.

Svejstrup, J.Q. (2007). Elongator complex: how many roles does it play? Curr. Opin. Cell Biol. *19*, 331–336.

Tammsalu, T., Matic, I., Jaffray, E.G., Ibrahim, A.F.M., Tatham, M.H., and Hay, R.T.

(2014). Proteome-Wide Identification of SUMO2 Modification Sites. Science Signaling *7*, rs2–rs2.

Tatham, M.H., Jaffray, E., Vaughan, O.A., Desterro, J.M., Botting, C.H., Naismith, J.H., and Hay, R.T. (2001). Polymeric chains of SUMO-2 and SUMO-3 are conjugated to protein substrates by SAE1/SAE2 and Ubc9. J Biol Chem *276*, 35368–35374.

Tatham, M.H., Geoffroy, M.-C., Shen, L., Plechanovová, A., Hattersley, N., Jaffray, E.G., Palvimo, J.J., and Hay, R.T. (2008). RNF4 is a poly-SUMO-specific E3 ubiquitin ligase required for arsenic-induced PML degradation. Nature *10*, 538–546.

Thakur, S.S., Geiger, T., Chatterjee, B., Bandilla, P., Fröhlich, F., Cox, J., and Mann, M. (2011). Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. Mol Cell Proteomics *10*, M110.003699.

Theile, C.S., Witte, M.D., Blom, A.E.M., Kundrat, L., Ploegh, H.L., and Guimaraes, C.P. (2013). Site-specific N-terminal labeling of proteins using sortase-mediated reactions. Nat Protoc *8*, 1800–1807.

Thomsen, M.C.F., and Nielsen, M. (2012). Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res *40*, W281–W287.

Ting, L., Cowley, M.J., Hoon, S.L., Guilhaus, M., Raftery, M.J., and Cavicchioli, R. (2009). Normalization and Statistical Analysis of Quantitative Proteomics Data Generated by Metabolic Labeling. Mcponline.org *8*, 2227–2242.

Ton-That, H., Labischinski, H., Berger-Bächi, B., and Schneewind, O. (1998). Anchor structure of staphylococcal surface proteins. III. Role of the FemA, FemB, and FemX factors in anchoring surface proteins to the bacterial cell wall. J Biol Chem *273*, 29143–29149.

Ton-That, H., Liu, G., Mazmanian, S.K., Faull, K.F., and Schneewind, O. (1999). Purification and characterization of sortase, the transpeptidase that cleaves surface proteins of Staphylococcus aureus at the LPXTG motif. Proc Natl Acad Sci USA *96*, 12424–12429.

Ton-That, H., and Schneewind, O. (2003). Assembly of pili on the surface of Corynebacterium diphtheriae. Mol Microbiol *50*, 1429–1438.

Truong, K., Lee, T.D., and Chen, Y. (2012). Small ubiquitin-like modifier (SUMO) modification of E1 Cys domain inhibits E1 Cys domain enzymatic activity. J Biol Chem *287*, 15154–15163.

Tsirigotis, M., Thurig, S., Dubé, M., Vanderhyden, B.C., Zhang, M., and Gray, D.A. (2001). Analysis of ubiquitination in vivo using a transgenic mouse model. BioTechniques *31*, 120–6–128–130.

Uzunova, K., Göttsche, K., Miteva, M., Weisshaar, S.R., Glanemann, C., Schnellhardt, M., Niessen, M., Scheel, H., Hofmann, K., Johnson, E.S., et al. (2007). Ubiquitin-dependent proteolytic control of SUMO conjugates. J Biol Chem *282*, 34167–34175.

Van der Veen, A.G., and Ploegh, H.L. (2012). Ubiquitin-like proteins. Annu. Rev.

Biochem. *81*, 323–357.

Van der Veen, A.G., Schorpp, K., Schlieker, C., Buti, L., Damon, J.R., Spooner, E., Ploegh, H.L., and Jentsch, S. (2011). Feature Article: Role of the ubiquitin-like protein Urm1 as a noncanonical lysine-directed protein modifier. Proc Natl Acad Sci USA.

Varshavsky, A. (2011). The N-end rule pathway and regulation by proteolysis. Protein Sci.

Vembar, S.S., and Brodsky, J.L. (2008). One step at a time: endoplasmic reticulum-associated degradation. Nat Rev Mol Cell Biol *9*, 944–957.

Vertegaal, A.C.O., Andersen, J.S., Ogg, S.C., Hay, R.T., Mann, M., and Lamond, A.I. (2006). Distinct and overlapping sets of SUMO-1 and SUMO-2 target proteins revealed by quantitative proteomics. Mol Cell Proteomics *5*, 2298–2310.

Wagner, S.A., Beli, P., Weinert, B.T., Nielsen, M.L., Cox, J., Mann, M., and Choudhary, C. (2011). A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. Mol Cell Proteomics *10*, M111.013284.

Wagner, S.A., Beli, P., Weinert, B.T., Schölz, C., Kelstrup, C.D., Young, C., Nielsen, M.L., Olsen, J.V., Brakebusch, C., and Choudhary, C. (2012). Proteomic analyses reveal divergent ubiquitylation site patterns in murine tissues. Mol Cell Proteomics *11*, 1578–1585.

Wang, D., and Cotter, R.J. (2005). Approach for determining protein ubiquitination sites by MALDI-TOF mass spectrometry. Anal Chem *77*, 1458–1466.

Wang, D., Xu, W., McGrath, S.C., Patterson, C., Neckers, L., and Cotter, R.J. (2005). Direct identification of ubiquitination sites on ubiquitin-conjugated CHIP using MALDI mass spectrometry. J Proteome Res *4*, 1554–1560.

Wang, J., Pérez-Santiago, J., Katz, J.E., Mallick, P., and Bandeira, N. (2010). Peptide identification from mixture tandem mass spectra. Mol Cell Proteomics *9*, 1476–1485.

Wang, X., Herr, R.A., and Hansen, T.H. (2012). Ubiquitination of substrates by esterification. Traffic *13*, 19–24.

Washburn, M.P., Wolters, D., and Yates, J.R. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol *19*, 242–247.

Wasinger, V.C., Zeng, M., and Yau, Y. (2013). Current status and advances in quantitative proteomic mass spectrometry. Int J Proteomics *2013*, 180605.

Wilkinson, K.A., and Henley, J.M. (2010). Mechanisms, regulation and consequences of protein SUMOylation. Biochem J *428*, 133–145.

Williamson, D.J., Fascione, M.A., Webb, M.E., and Turnbull, W.B. (2012). Efficient N-terminal labeling of proteins by use of sortase. Angew Chem Int Ed Engl *51*, 9377–9380.

Williamson, D.J., Webb, M.E., and Turnbull, W.B. (2014). Depsipeptide substrates for sortase-mediated N-terminal protein ligation. Nat Protoc *9*, 253–262.

Winkler, G.S., Kristjuhan, A., Erdjument-Bromage, H., Tempst, P., and Svejstrup, J.Q. (2002). Elongator is a histone H3 and H4 acetyltransferase important for normal histone acetylation levels in vivo. Proc Natl Acad Sci USA *99*, 3517–3522.

Wohlschlegel, J.A., Johnson, E.S., Reed, S.I., and Yates, J.R. (2006). Improved identification of SUMO attachment sites using C-terminal SUMO mutants and tailored protease digestion strategies. J Proteome Res *5*, 761–770.

Wong, C.C.L., Cociorva, D., Venable, J.D., Xu, T., and Yates, J.R. (2009). Comparison of different signal thresholds on data dependent sampling in Orbitrap and LTQ mass spectrometry for the identification of peptides and proteins in complex mixtures. J Am Soc Mass Spectrom *20*, 1405–1414.

Wu, Z., Guo, X., and Guo, Z. (2011). Sortase A-catalyzed peptide cyclization for the synthesis of macrocyclic peptides and glycopeptides. Chem. Commun. *47*, 9218–9220.

Xi, J., Ge, Y., Kinsland, C., McLafferty, F.W., and Begley, T.P. (2001). Biosynthesis of the thiazole moiety of thiamin in Escherichia coli: identification of an acyldisulfide-linked protein--protein conjugate that is functionally analogous to the ubiquitin/E1 complex. Proc Natl Acad Sci USA *98*, 8513–8518.

Xie, Y., Kerscher, O., Kroetz, M.B., McConchie, H.F., Sung, P., and Hochstrasser, M. (2007). The yeast Hex3.Slx8 heterodimer is a ubiquitin ligase stimulated by substrate sumoylation. J Biol Chem *282*, 34176–34184.

Xu, G., Paige, J.S., and Jaffrey, S.R. (2010). Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. Nat Biotechnol *28*, 868–873.

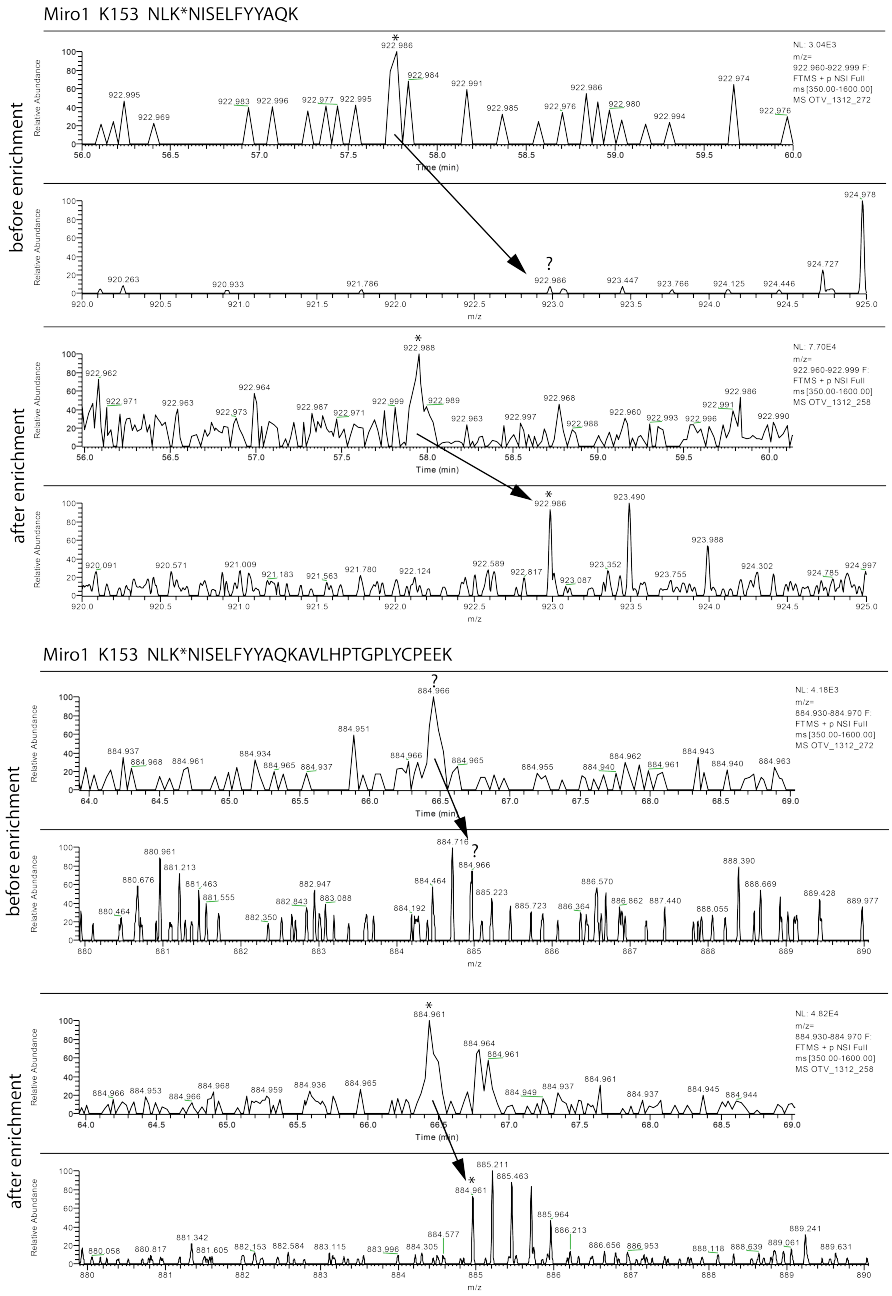
Xu, P., Duong, D.M., Seyfried, N.T., Cheng, D., Xie, Y., Robert, J., Rush, J., Hochstrasser, M., Finley, D., and Peng, J. (2009). Quantitative proteomics reveals the function of unconventional ubiquitin chains in proteasomal degradation. Cell *137*, 133–145.

Zhu, G., Sun, L., Yan, X., and Dovichi, N.J. (2013). Single-shot proteomics using capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry with production of more than 1250 Escherichia coli peptide identifications in a 50 min separation. Anal Chem *85*, 2569–2573.

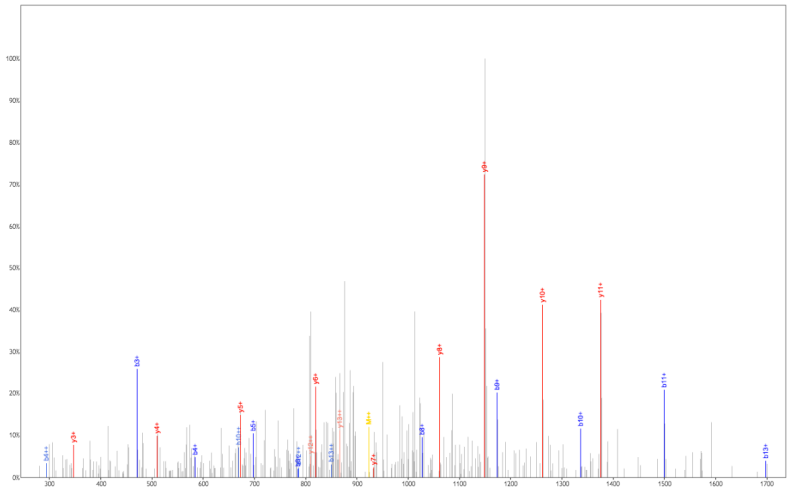
Zhu, S., Goeres, J., Sixt, K.M., Békés, M., Zhang, X.-D., Salvesen, G.S., and Matunis, M.J. (2009). Protection from isopeptidase-mediated deconjugation regulates paralog-selective sumoylation of RanGAP1. Mol Cell *33*, 570–580.

Appendix A:
Miro1 isopeptides

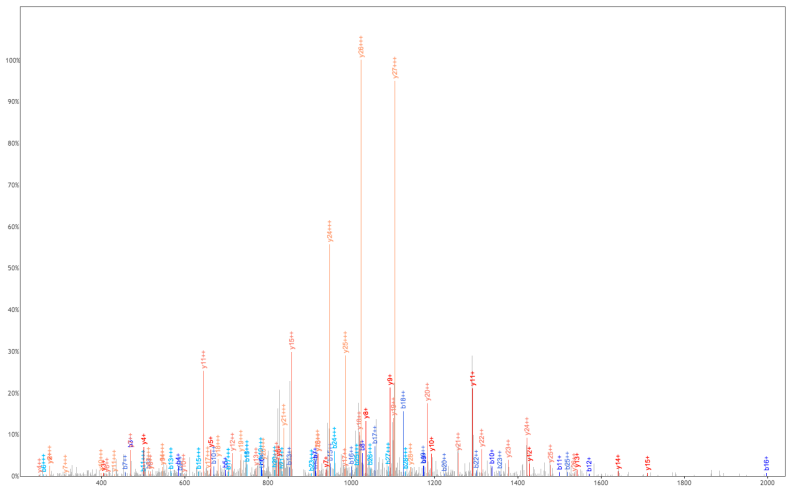
Evidence for additional Miro1 K153 and K406 isopepetides. K572 is presented in the main text of chapter 2.



Extracted ion chromatograms are indicated (*, or ? for uncertain or absent signal) for Miro1 K153 isopeptides NLK(GG)NISELFYYAQK (922.98m/z, 2+) and NLK(GG)NISELFYYAQKAVLHPTGPLYCPEEK (884.95, 4+) before and after isopeptide

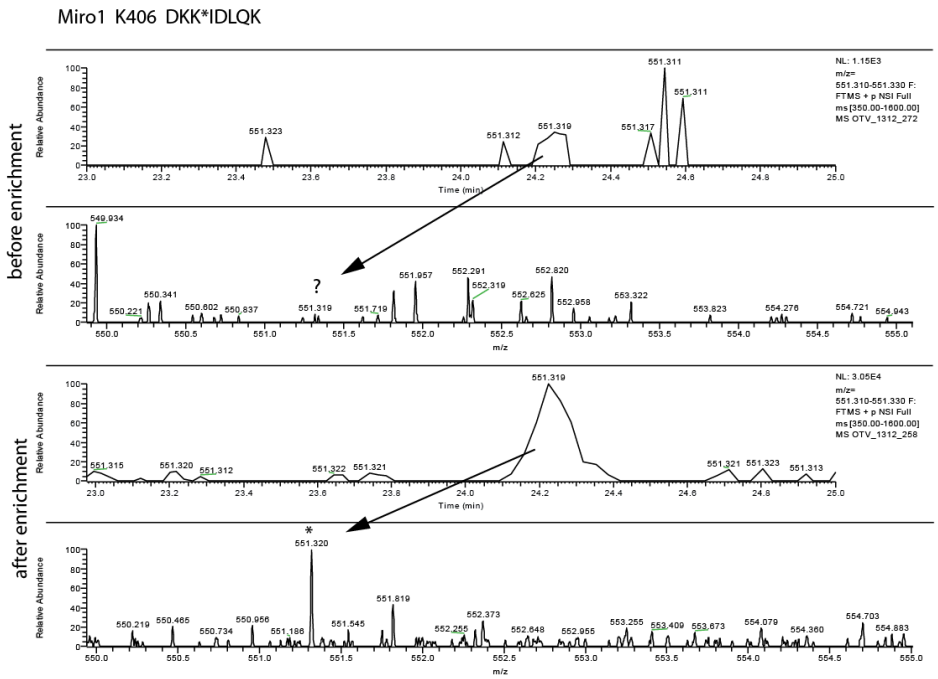


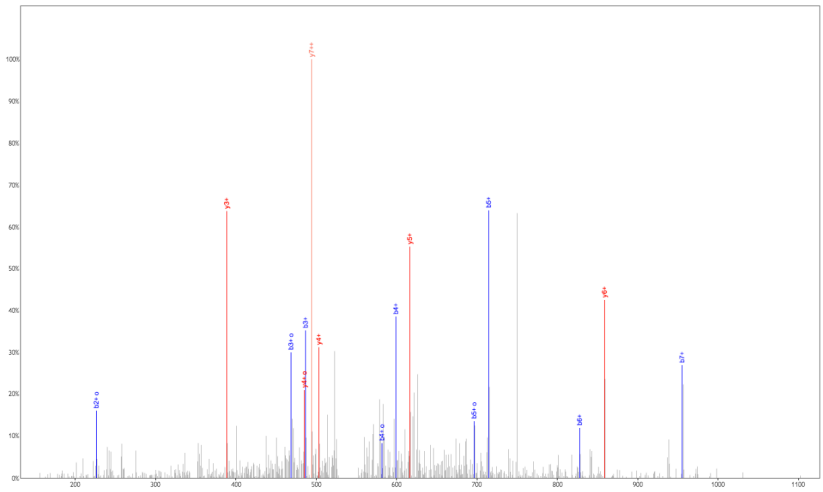
115.0502	58.0287	1	N	14		
228.1343	114.5708	2	L	13	1730.9112	865.9592
470.2722	235.6397	3	K	12	1617.8271	809.4172
584.3151	292.6612	4	N	11	1375.6892	688.3483
697.3992	349.2032	5	I	10	1261.6463	631.3268
784.4312	392.7192	6	S	9	1148.5623	574.7848
913.4738	457.2405	7	E	8	1061.5302	531.2687
1026.5578	513.7826	8	L	7	932.4876	466.7475
1173.6262	587.3168	9	F	6	819.4036	410.2054
1336.6896	668.8484	10	Y	5	672.3352	336.6712
1499.7529	750.3801	11	Y	4	509.2718	255.1395
1570.7900	785.8986	12	A	3	346.2085	173.6079
1698.8486	849.9279	13	Q	2	275.1714	138.0893
		14	K	1	147.1128	74.0600



b+	b2+	b3+	#	Seq	#	y+	y2+	y3+
115.1099	58.0586	39.0415	1	N	29			
228.2676	114.6374	76.7607	2	L	28	3424.8063	1712.9068	1142.2736
470.4828	235.7450	157.4991	3	K	27	3311.6486	1656.3280	1104.5544
584.5854	292.7963	195.5333	4	N	26	3069.4334	1535.2204	1023.8160
697.7430	349.3752	233.2525	5	I	25	2955.3308	1478.1690	985.7818
784.8203	392.9138	262.2783	6	S	24	2842.1732	1421.5902	948.0626
913.9343	457.4708	305.3163	7	E	23	2755.0959	1378.0516	919.0368
1027.0920	514.0496	343.0355	8	L	22	2625.9819	1313.4946	875.9988
1174.2658	587.6366	392.0935	9	F	21	2512.8242	1256.9158	838.2796
1337.4391	669.2232	446.4845	10	Y	20	2365.6504	1183.3288	789.2216
1500.6123	750.8098	500.8756	11	Y	19	2202.4771	1101.7422	734.8306
1571.6902	786.3488	524.5683	12	A	18	2039.3039	1020.1556	680.4395
1699.8195	850.4134	567.2780	13	Q	17	1968.2260	984.6166	656.7468
1827.9917	914.4995	610.0021	14	K	16	1840.0967	920.5520	614.0371
1899.0696	950.0385	633.6947	15	A	15	1711.9245	856.4659	571.3130
1998.2007	999.6040	666.7384	16	V	14	1640.8466	820.9269	547.6204
2111.3583	1056.1828	704.4576	17	L	13	1541.7155	771.3614	514.5767
2248.4976	1124.7524	750.1707	18	H	12	1428.5579	714.7826	476.8575
2345.6128	1173.3100	782.5425	19	P	11	1291.4186	646.2129	431.1444
2446.7167	1223.8620	816.2437	20	T	10	1194.3034	597.6553	398.7726
2503.7680	1252.3876	835.2609	21	G	9	1093.1995	547.1034	365.0714
2600.8832	1300.9452	867.6326	22	P	8	1036.1482	518.5777	346.0542
2714.0408	1357.5240	905.3518	23	L	7	939.0330	470.0201	313.6825
2877.2141	1439.1107	959.7429	24	Y	6	825.8754	413.4413	275.9633
3037.3784	1519.1929	1013.1310	25	C	5	662.7021	331.8547	221.5722
3134.4936	1567.7505	1045.5027	26	P	4	502.5377	251.7725	168.1841
3263.6076	1632.3074	1088.5407	27	E	3	405.4226	203.2149	135.8124
3392.7216	1696.8644	1131.5787	28	E	2	276.3086	138.6579	92.7744
			29	K	1	147.1946	74.1009	49.7364

MS2 spectral support for identification of K153 isopeptides NLK(GG)NISELFYYAQK (922.98m/z, 2+) and NLK(GG)NISELFYYAQKAVLHPTGPLYCPEEK (884.95, 4+)





b+	#	Seq	#	y+	y2+
116.0342	1	D	8		
244.1292	2	K	7	986.5993	493.8033
486.2671	3	K	6	858.5043	429.7558
599.3511	4	I	5	616.3665	308.6869
714.3781	5	D	4	503.2824	252.1448
827.4621	6	L	3	388.2554	194.6314
955.5207	7	Q	2	275.1714	138.0893
	8	K	1	147.1128	74.0600

Extracted ion chromatograms are indicated (*, or ? for uncertain or absent signal) for Miro1 K406 isopeptides DKK(GG)IDLQK (551.32m/z, 2+) before and after enrichment with MS2 spectral support.

Appendix B:
ICID use details

Execution on mac/linux terminal:

```
cd /usr/bin/java/VKJLib/bin
java -Xmx1024m -classpath .:jopt-simple-4.3.jar:jrap_StAX_v5.2.jar
PeakPicker.PeakPicker -p paramfile -m mzXMLfile -f fastaDBfile >ouput.txt
```

Parameter file template, with comments (currently set for tryptic SUMO2):

```
#the isotopic difference between label(s). Must be a space delimited list of masses
if more than one label expected.
#Examples:
#mTRAQ 4.0070994 8.0141988 16.0283976
#N15sumo2/3 = 37.88733 tryp, 5.98220982 for chymo QQQTGG
#heavy-Phe: 13C=+1.00335483, 13C6=6.020129, 2*13C6F=12.040258, 3*13C6=18.060387
#     due to missed cleavage you will need to do both 12.040258 and 18.060387
#heavy-Tyr: 13C9 = 9.03019347
deltaMasses = 12.040258 18.060387

#the modifying mass(es) of the light isotope label(s).  Must be one for each
deltaMasses entry above.
#Must be a space delimited list of masses.
#t_SUMO2/3 (FR)FDGQPINETDTPAQLEMEDEDTIDVFQQQTGG 3549.53652, 3565.531435 Mox,
3852.70604 1mc, 3868.70094 Mox+1mc
#Only need one per isotope label though (ie Mox does not change labelling)
#eg 140.094963 for mTRAQ
#599.266339 for chymo_SUMO2/3 C-term pep
#0 for silac
#STLHLVLRLRGG = 1302.78841, ESTLHLVLRLRGG = 1431.83100
labelMasses = 3549.53652 3852.70604


######### parameters for detecting isotope coded feature pairs
# inter-spectrum mass error - used between spectra while finding features (ideally
~2*instrument error. The mass difference between peaks could be: m/z+err - m/z-err =
2*err)
ppmError = 10.0

# intra-spectrum ppm error to use during isotope pair matching (in-spectrum will have
a higher mass accuracy than between spectra)
ppmErrorDeltaMass = 2.0

# ppm error used when comparing masses within isotope clusters (ie compare
monoisotope and second isotope within the same spectrum)
ppmErrorIsotopes = 2.0

# feature apex must be above this threshold
intensityCutoff = 10000

#min and max charge states to consider
minZ = 3
maxZ = 9

#chromatographic time deviation to consider. Units = mins.metric (ie 0.05=3seconds),
the time permitted between apices to be considered as a co-eluting isotope pair
maxTimeDeviation = 0.05
```

```
# these values are log2(L/H) (eg -1 -> 1 is a 2-fold tolerance) - permits non-
symmetric ratios if labels were not mixed at 50:50
featureRatioMinLog2 = -1
featureRatioMaxLog2 = 1

# true permits feature recognition from only 2 peaks, otherwise 3 or more points are
required for an apex (only use true if you want to really dig into the noise).
permitDoubletFeature = false

# if true, small interference surrounding a major peak are removed.  Done at the peak
picking stage (before finding feature pairs. Satellite params only take effect if
# will look +/-ppmSatellites for a much larger peak (the real peak) and remove this
peak if less than percentSatellites relative to the major/real peak
removeSatellitePeaks is true.
removeSatellitePeaks = true
ppmSatellites = 80
percentSatellites = 2.5

############ params for filtering noise out of the feature-pair list
# Removes isotope feature pairs where another isotope feature pair exists at the
expected lower isotope mass (only checks within feature set)
# what does this first param do?
deisotopeFeatureSet = false
# remove isotope features where the low feature is clearly not the monoisotopic peak
deisotopeLowMass = true
# remove isotope features where the high feature is clearly not the monoisotopic peak
(should not do if impure N15/silac)
deisotopeHighMass = false
# mins from the peak apex at which a lower intensity isotope pair will be removed
(only keep the most intense feature pair)
chromatographicPeakTolerance = 1.5


########## parameters regarding what to output to the inclusion list
# mins.metric elution time will be centered at inclListTimeWindow/2
inclListTimeWindow = 3
# at least one of includeLowMass or includeHighMass must be true
includeLowMass = true
includeHighMass = false
# minimum intensity for a feature to be output to inclusion list. Will only have an
effect if > intensityCutOff
includeFromIntensity = 0
# maximum intensity for a feature to be output to inclusion list.  Use a very large
number (eg 1.7976931348623157E308 = Double.MAX_VALUE = default) for all peaks
includeToIntensity = 1.7976931348623157E308

# info required to digest the fasta file into peptides. DB must be given as separate
terminal/cmd param (enzyme is assumed to be trypsin)
ppmPeptide = 10.0
```

To compile SUMmOn results with ICID results, a list of summon directories are required, one per line in a text file
ICID results are optional if you just want to consolidate SUMmOn search results

Execution on mac/linux terminal:

```
cd /usr/bin/java/VKJLib/bin
java -cp .:jopt-simple-4.3.jar vk.proteomics.summon.SummonResultSet -s
inputSummonDirectoryListFile -o outputResultsFile -i ICIDresultsFile
```

```
Parameters (with defaults assigned):
numIsotopes = 5
isotopeDelta = 1.003355
ppmErrorTolerance = 10.0
useModHyperlinks = true
useCandHyperlinks = false
minModScore = 0.8
minCandidateScoreWithAccurateMass = 0.5
minCandidateScoreWithoutAccurateMass = 999.0
minOrphanModScore = 999.0
scanTolerance = 250
```

Appendix C:
Calibrator use details

Calibrator parameters:

Usage:	java -cp .:jopt-simple-4.3.jar:commons-codec-1.6.jar:commons-math3-3.0.jar:jrap_StAX_v5.2.jar LCMSrecalibration.Calibrator params
-m string	path to mzXML input file (centroid data)
-p string	path to pepXML input file - MS1 spectra will be recalibrated using peptide IDs (no recalibration occurs without this option)
-v double	minimum p-value required to import an ID from the pepXML input file (optional, accepts 0-1, default = 1.0, only valid if using option p)
-i double	minimum intensity of ms2 precursor ion required to import ID (default = 0.0)
-s [int]	minimum data points required to do a single-scan calibration (only valid if using option f; Default = 25; no value turns single scan calibration off)
-d [string]	decoy substring as part of protein accession to filter out reverse/random DB hits (no param turns filtering off; default string = rev_)
-x m|t	which axis to calibrate. m = MZ axis, t = time axis. (Default = both MZ and time axes)
-l double	lockmass m/z. Only use spectra with a lockmass for mz calibration and correct ppm before calibration
-L double	lockmass m/z. Only use spectra with a lockmass for mz calibration but DO NOT correct ppm before calibration
-T double	lockmass ppm tolerance. Only required if using l or L (default = 10.0 ppm)
-t double	ppm tolerance used during precursor detection and correction (optional; default = 25.0 ppm)
-f	use features from label-free XPress quant to calibrate with precursor m/z over the chromatographic elutions. Requires XPress LF quant in pepxml. (Default=false)
-c	correct MS2 precursor m/z values from observed MS1 spectra (useful if ms2 precursors were determined from preview scans)
-z [double]	zero MS2 precursor m/z if no ms1 precursor observed above the specified intensity (default=0), and do not use in calibration (only valid if using -c)

-r	restrict recalibration to IDs with corrected precursors (only valid if using options p and c)
-a	average the ppm from multiple data points from a single scan (default = natural-log-intensity weighted PPM average)
-w	suppress writing of mzXML file (useful if you just want to see the stats output to optimise settings)
-h	print this help info
--atl	alignment of time, max number of Lines (default = 10000)
--atd	alignment of time, min number of Data points per subset (default = 100)
--ats	alignment of time, remove outliers if outside n Standard deviations (default = 4.0)
--aml	alignment of m/z, max number of Lines (default = 50)
--amd	alignment of m/z, min number of Data points per subset (default = 200)
--ams	alignment of m/z, remove outliers if outside n Standard deviations (default = 3.0)
Example:	-crzfa -v 0.95 -s=20 -i5000 -t 15.0 -drev_ -l 519.138816 -T 10.0 -m 'mzXMLfilePath' -p 'pepXMLfilePath'

For convenience, use a bash script called ‘jCalibrator.sh’ to manage classpath details

```
#!/usr/bin/env bash
# To execute, just call
#          jCalibrator.sh [params]
cd /usr/bin/java/VKJLib/bin/
java -cp .:../../jopt-simple-4.3.jar:../../commons-codec-1.6.jar:../../commons-math3-
3.0.jar:../../jrap_StAX_v5.2.jar LCMSrecalibration.Calibrator $*
```

Appendix D:
HyperProphet use details

HyperProphet parameters:

Usage:	java -cp .:jopt-simple-4.3.jar:commons-math3-3.0.jar:jrap_StAX_v5.2.jar HyperProphet.HyperProphet params
-d string	data directory (containing *.pep.xml and corresponding *.mzXML files) [required]
-i double	input min p-val (peptide ID p-val, eg 1%FDR to prevent unnecessary ID transfer) [optional; default = -o output]
-a double	alignment min p-val (used for time alignment) [optional; default = 0.99]
-o double	output min p-val (for assignment to other aligned runs). Must be >= input p-val. [optional; default = 0.85]
-m double	minTimeSecs [optional; default = 300]
-x double	maxTimeSecs [optional; default = a very big number = the end of the run]
-s string	subset to output new mzXML files for, as a comma separated list without spaces [optional; default = all files]. Do not include both light and heavy single channel runs that cannot be aligned together
-l	align using MSn time where LabelFreeQuant unavailable [optional; default = only use label-free apex time]. This option will include IDs with missing quant. Also permits processing files without having done XPressLableFreeQuant (not recommended).
e [double]	do not export using MSn time where LabelFreeQuant unavailable [optional; default = do export using MSn time if LabelFreeQuant apex time unavailable]. If the value is specified, features must also have an apex greater than this intensity (default = 0, ie all spectra with detected features)
-g	use a segmentedAligner for chromatographic alignment [optional; default = use averagingAligner which enforces contiguous best fit lines]
--al int	Align with max Lines [optional, default = 50]
--ad int	Align with min Data points per line [optional, default = 200]
--as double	Align with removal of outliers if ouside this Stdev [optional, default = 2]
-p double	set max ppm tolerance (+/-) for peptide IDs to be transfered [optional; default = 5ppm]

-T double	set max time tolerance (+/-seconds) for seeking precursor/feature in target extracted ion chromatogram traces. Seeking better transfer time turned off if set to zero [optional; default = 30]
-p double	set max ppm tolerance (+/-) for extracting ion chromatogram traces [optional; only used if t>0; default = 10]
-i double	set intensity threshold for target extracted ion chromatogram traces. Any trace <= to this will not be transferred [optional; default = 1000]
-c	suppress output of the file containing extracted ion Chromatogram file containing traces and MS2 tansfer points [optional; default = true]
-b	set m/z to theoretical for seeking precursor/feature in target extracted ion chromatogram traces [optional; default = observed]
-h	print this help, which is pretty useless since you're already reading this

For convenience, use a bash script called 'jHyperProphet.sh' to manage classpath details

```
#!/usr/bin/env bash
# To execute, just call
#          jHyperProphet.sh [params]
cd /usr/bin/java/VKJLib/bin
java -cp .:jopt-simple-4.3.jar:commons-math3-3.0.jar:jrap_StAX_v5.2.jar
HyperProphet.HyperProphet $*
```

Appendix D:

Gene ontology results for down regulated proteins in elp3/wt at 1%FDR (followed by list of proteins)

funspec: P-value cutoff=0.01 with Bonferroni correction				
GO Molecular Function (1646 categories)				
Category	**p-value**	**In Category from Cluster**	**k**	**f**
nucleotide binding [GO:0000166]	9.18E-11	CDC19 ADE1 MIS1 TEF2 GRS1 CDC28 SSE2 ISW1 CDC10 PGK1 DBP10 PSA1 VMA1 NOP6 SHS1 SES1 RLI1 SRP101 YDR341C YRA1 EFT2 HPT1 TIF35 SAM2 GIN4 EMI2 GCD11 ARB1 MOT2 ACT1 PMA1 LSG1 INO80 HXK2 ADE6 TYS1 ADE3 PFK1 YGR250C STE20 SBP1 GPA1 DED81 YHR020W CDC12 YCK1 TIM44 THS1 CCT8 YJL055W GSH1 CCT7 MET3 PTK2 ADO1 MET14 URA6 YNK1 LHS1 OXP1 VPS1 ACS2 CDC42 RCK2 YEF3 MYO5 PFK2 HRB1 PUB1 NOP15 NOP13 RHO5 SSB2 YNL247W NRD1 MCK1 ACC1 NOP12 RFC4 EFT1 RFC1 MYO2 PRT1 NOG1 SSE1 CDC60 CBC2 RVB2 GLN1 TEF1 RHO1	91	778
RNA binding [GO:0003723]	3.39E-08	NCL1 LSM2 RPS11B PAT1 DBP10 TRM8 NOP6 RPS11A MRPL1 TMA64 YRA1 EFT2 TIF35 THO1 MOT2 LSM4 TIF4632 KEM1 TAN1 TIF4631 YGR250C SBP1 LSM12 SCP160 LSM1 SUI2 TMA22 CBF5 YEF3 RPL6B HRB1 PUB1 NOP15 LSM7 NOP13 NRD1 PUS4 BRE5 NOP12 EFT1 BFR1 PRT1 RPS9A NOP53 CBC2 PUS1 SRP68	47	337
translation initiation factor activity [GO:0003743]	1.55E-07	RPG1 RLI1 TMA64 TIF35 GCD11 TIF4632 YGR054W TIF4631 SUI2 TMA22 TIF34 NIP1 PRT1	13	38
tubulin binding [GO:0015631]	3.08E-07	NUM1 GIM4 PAC10 GIM5 GIM3 RBL2	6	7

GO Biological Process (2062 categories)				
Category	**p-value**	**In Category from Cluster**	**k**	**f**
translational initiation [GO:0006413]	1.24E-08	RPG1 RLI1 TMA64 TIF35 GCD11 TIF4632 YGR054W TIF4631 SUI2 TMA22 CLU1 TIF34 NIP1 SIS1 PRT1	15	43
glycolysis [GO:0006096]	2.81E-07	CDC19 PGI1 PGK1 EMI2 HXK2 PFK1 KGD1 TDH2 PFK2 GPM3 YOR283W	11	28
translation [GO:0006412]	2.25E-06	RPS11B TRM7 RPG1 TEF2 GRS1 RPS6B SES1 RPS11A RLI1 MRPL1 YDR341C EFT2 TIF35 GCD11 RPL22B TIF4632 YGR054W TIF4631 TYS1 RPL8A DED81 YHR020W THS1 RSM25 SUI2 YEF3 RPL6B TIF34 NIP1 RPL9B SSB2 YNL247W BRX1 RPL3 TMA46 EFT1 PRT1 RPS9A RPS6A CDC60 TEF1	41	318
'de novo' IMP biosynthetic process [GO:0006189]	3.32E-06	ADE1 ADE8 ADE6 ADE13 ADE17 ADE4	6	9
ribosomal subunit export from nucleus [GO:0000054]	3.43E-06	RLI1 LSG1 ZUO1 RNA1 SSB2 NOC2 NOG1	7	13
purine nucleotide biosynthetic process [GO:0006164]	4.16E-06	ADE1 MIS1 ADE8 ADE6 ADE3 ADE13 ADE17 ADE4	8	18
tubulin complex assembly [GO:0007021]	4.43E-06	GIM4 PAC10 GIM5 GIM3 RBL2	5	6

List of all proteins down regulated in elp3/wt at 1%FDR

Systematic Name	Standard Name	logFC	adj.P.Val
YBR093C	Pho5p	-2.4642	3.0500E-06
YNL160W	Ygp1p	-2.3280	7.2523E-04
R0040C	Rep2p	-1.4263	7.0000E-06
YPL199C		-1.1242	2.4100E-06
YIL063C	Yrb2p	-1.1084	2.7600E-05
YMR278W	Prm15p	-1.0254	2.6338E-04
YML123C	Pho84p	-1.0188	9.7700E-05
YNL251C	Nrd1p	-0.9419	7.1334E-04
YCR005C	Cit2p	-0.9092	2.8100E-05
YDL147W	Rpn5p	-0.8990	2.9285E-04
YHR087W	Rtc3p	-0.8959	2.4626E-03
YKL157W	Ape2p	-0.8844	4.0200E-05

YKL001C	Met14p	-0.8808	1.4261E-04
YBR169C	Sse2p	-0.8782	3.7934E-04
YAR042W	Swh1p	-0.8550	1.5251E-04
YLR354C	Tal1p	-0.8436	1.1100E-05
YKR016W	Fcj1p	-0.8375	1.9673E-04
YLR002C	Noc3p	-0.8329	1.1455E-04
YMR205C	Pfk2p	-0.8171	7.0000E-06
YDL213C	Nop6p	-0.7964	1.0067E-04
YPL019C	Vtc3p	-0.7961	1.6700E-05
YOR206W	Noc2p	-0.7764	3.6700E-05
YJL088W	Arg3p	-0.7762	1.5786E-03
YDR507C	Gin4p	-0.7754	7.1334E-04
YMR120C	Ade17p	-0.7741	1.9359E-04
YPR072W	Not5p	-0.7662	3.2400E-05
YFR030W	Met10p	-0.7658	2.7600E-05
YNL180C	Rho5p	-0.7618	7.8600E-05
YGR233C	Pho81p	-0.7410	2.7600E-05
YLR206W	Ent2p	-0.7335	5.3256E-04
YKL065C	Yet1p	-0.7312	5.7800E-06
YHR107C	Cdc12p	-0.7234	1.6086E-04
YJR007W	Sui2p	-0.7098	1.4500E-05
YDR233C	Rtn1p	-0.7054	7.0000E-06
YIL015W	Bar1p	-0.7027	7.8900E-06
YOR198C	Bfr1p	-0.6952	4.9900E-06
YLR248W	Rck2p	-0.6794	3.8206E-04
YFL034C-A	Rpl22Bp	-0.6730	7.0000E-06
YOR163W	Ddp1p	-0.6708	7.0700E-06
YGL232W	Tan1p	-0.6571	4.5700E-05
YDR117C	Tma64p	-0.6482	5.1400E-05
YNL231C	Pdr16p	-0.6454	1.6900E-05
YMR275C	Bul1p	-0.6448	5.9226E-03
YJR137C	Met5p	-0.6408	7.6300E-06
YER044C	Erg28p	-0.6401	5.0900E-05
YGL173C	Xrn1p	-0.6392	7.4800E-06
YPL081W	Rps9Ap	-0.6382	2.9300E-05
YMR047C	Nup116p	-0.6362	7.9055E-04
YML094W	Gim5p	-0.6323	2.9400E-05
YNL168C	Fmp41p	-0.6238	3.4000E-05
YHR005C	Gpa1p	-0.6238	5.6800E-05

YMR072W	Abf2p	-0.6230	1.7000E-05
YHR201C	Ppx1p	-0.6202	2.7600E-05
YEL071W	Dld3p	-0.6190	1.0500E-05
YBR092C	Pho3p	-0.6113	2.7600E-05
YFL045C	Sec53p	-0.6104	7.0000E-06
YGL150C	Ino80p	-0.6062	3.5000E-05
YHR113W	Ape4p	-0.6016	1.4696E-04
YLR003C	Cms1p	-0.6014	5.7960E-04
YFR041C	Erj5p	-0.5984	1.0200E-05
YLR449W	Fpr4p	-0.5947	2.8300E-05
YER091C	Met6p	-0.5931	1.8400E-05
YHR135C	Yck1p	-0.5840	1.6927E-03
YLR418C	Cdc73p	-0.5829	4.3100E-05
YDL201W	Trm8p	-0.5811	3.1600E-05
YGL253W	Hxk2p	-0.5750	3.5423E-04
YNL134C		-0.5749	1.5977E-03
YDR481C	Pho8p	-0.5733	5.1300E-05
YBR159W	Ifa38p	-0.5681	3.8100E-05
YHR097C		-0.5671	7.3903E-03
YNR016C	Acc1p	-0.5667	6.3900E-05
YPR032W	Sro7p	-0.5647	1.0372E-04
YAR015W	Ade1p	-0.5626	1.3000E-05
YLR229C	Cdc42p	-0.5611	8.7100E-05
YJL124C	Lsm1p	-0.5577	9.9100E-06
YGR082W	Tom20p	-0.5540	1.7406E-03
YLR153C	Acs2p	-0.5531	7.0700E-06
YGR019W	Uga1p	-0.5483	1.6203E-03
YGL206C	Chc1p	-0.5477	7.2300E-06
YOL012C	Htz1p	-0.5434	5.8100E-05
YOR054C	Vhs3p	-0.5428	1.1903E-03
YDR150W	Num1p	-0.5352	5.2567E-03
YER043C	Sah1p	-0.5316	3.6700E-05
YOR116C	Rpo31p	-0.5301	2.2692E-04
YLR249W	Yef3p	-0.5280	1.8810E-04
YIL022W	Tim44p	-0.5261	3.3200E-05
YEL003W	Gim4p	-0.5247	1.0232E-04
YPR060C	Aro7p	-0.5179	1.0623E-03
YDL074C	Bre1p	-0.5136	8.9838E-03
YJR010W	Met3p	-0.5129	1.5490E-04

YKL179C	Coy1p	-0.5123	5.2244E-04
YMR173W	Ddr48p	-0.5100	9.5900E-05
YGR286C	Bio2p	-0.5074	4.8348E-03
YMR099C		-0.5048	2.7200E-05
YDR153C	Ent5p	-0.5021	7.5600E-05
YLR309C	Imh1p	-0.5019	1.0459E-03
YJR051W	Osm1p	-0.4940	6.9919E-04
YDR170C	Sec7p	-0.4929	8.7300E-05
YPL231W	Fas2p	-0.4905	2.4900E-05
YML017W	Psp2p	-0.4890	1.4696E-04
YML074C	Fpr3p	-0.4890	3.6000E-05
YDR098C	Grx3p	-0.4886	3.2489E-03
YKR003W	Osh6p	-0.4829	5.6579E-04
YMR307W	Gas1p	-0.4792	1.0498E-04
YKL104C	Gfa1p	-0.4787	7.3900E-05
YDR091C	Rli1p	-0.4766	1.4666E-04
YPL210C	Srp72p	-0.4764	8.3401E-04
YHR049W	Fsh1p	-0.4748	6.4400E-05
YKL182W	Fas1p	-0.4740	1.6100E-05
YDR518W	Eug1p	-0.4703	8.0845E-04
YGL049C	Tif4632p	-0.4672	1.3325E-04
YNL147W	Lsm7p	-0.4650	2.2764E-03
YPL146C	Nop53p	-0.4644	2.7609E-03
YJL012C	Vtc4p	-0.4643	3.1000E-05
YBR222C	Pcs60p	-0.4638	1.0453E-04
YEL015W	Edc3p	-0.4636	3.1100E-04
YNL141W	Aah1p	-0.4624	5.1243E-04
YDR516C	Emi2p	-0.4605	7.2176E-04
YAL023C	Pmt2p	-0.4600	2.0627E-04
YJL053W	Pep8p	-0.4573	6.3400E-05
YGR009C	Sec9p	-0.4551	1.7325E-04
YOR265W	Rbl2p	-0.4538	2.7400E-05
YMR012W	Clu1p	-0.4528	5.1086E-04
YJL080C	Scp160p	-0.4524	1.3000E-05
YLR175W	Cbf5p	-0.4511	2.2100E-05
YIL138C	Tpm2p	-0.4497	9.0661E-04
YGL153W	Pex14p	-0.4496	3.8100E-05
YMR235C	Rna1p	-0.4444	4.0141E-03
YLR033W	Rsc58p	-0.4344	8.2481E-04

YHR039C	Msc7p	-0.4317	1.9815E-03
YER112W	Lsm4p	-0.4276	3.6700E-05
YDR023W	Ses1p	-0.4263	3.6700E-05
YBR109C	Cmd1p	-0.4261	1.0397E-04
YHR019C	Ded81p	-0.4255	3.6700E-05
YJL167W	Erg20p	-0.4235	6.5300E-05
YDR502C	Sam2p	-0.4226	3.6500E-05
YIL038C	Not3p	-0.4149	1.2481E-04
YIL078W	Ths1p	-0.4139	2.1900E-05
YER110C	Kap123p	-0.4125	3.4000E-05
YMR016C	Sok2p	-0.4123	4.2231E-04
YJR117W	Ste24p	-0.4073	3.8152E-04
YDL052C	Slc1p	-0.4068	1.2781E-04
YHR020W		-0.4065	8.1458E-04
YDR408C	Ade8p	-0.4059	8.6969E-03
YNL004W	Hrb1p	-0.4041	2.1185E-03
YDL225W	Shs1p	-0.4027	3.5002E-03
YCR016W		-0.4022	6.5077E-03
YDL130W-A	Stf1p	-0.4017	2.0680E-03
YBR234C	Arc40p	-0.4000	2.1171E-04
YGR054W		-0.3999	4.3800E-05
YMR027W		-0.3997	1.9243E-03
YDR292C	Srp101p	-0.3990	7.2316E-03
YLR095C	Ioc2p	-0.3975	2.9410E-04
YBR245C	Isw1p	-0.3942	3.7600E-05
YFL004W	Vtc2p	-0.3940	2.2494E-03
YKL073W	Lhs1p	-0.3925	1.4652E-03
YER068W	Mot2p	-0.3922	5.5924E-04
YDR145W	Taf12p	-0.3920	8.3500E-05
YEL042W	Gda1p	-0.3917	6.1783E-03
YDL145C	Cop1p	-0.3910	2.5121E-04
YGR240C	Pfk1p	-0.3909	2.9300E-05
YNL313C	Emw1p	-0.3904	1.4065E-04
YGR061C	Ade6p	-0.3877	3.5136E-04
YKL067W	Ynk1p	-0.3866	2.2683E-03
YGR185C	Tys1p	-0.3824	3.1000E-05
YNL175C	Nop13p	-0.3821	6.7600E-05
YGR250C		-0.3813	2.1785E-03
YLR045C	Stu2p	-0.3809	2.0413E-03

YCR030C	Syp1p	-0.3769	1.9873E-03
YBR160W	Cdc28p	-0.3704	4.8545E-04
YCR012W	Pgk1p	-0.3701	2.8300E-05
YPR080W	Tef1p	-0.3687	5.8100E-05
YBR118W	Tef2p	-0.3687	5.8100E-05
YBR079C	Rpg1p	-0.3677	3.1000E-05
YJR009C	Tdh2p	-0.3667	5.0900E-05
YPL094C	Sec62p	-0.3650	1.0764E-04
YDR399W	Hpt1p	-0.3632	5.1300E-05
YAR007C	Rfa1p	-0.3617	4.9380E-04
YPR108W	Rpn7p	-0.3607	3.8100E-05
YBR281C	Dug2p	-0.3582	1.0809E-03
YNL138W	Srv2p	-0.3572	1.2986E-03
YER063W	Tho1p	-0.3568	1.9890E-04
YDL099W	Bug1p	-0.3565	2.0513E-04
YFL010C	Wwm1p	-0.3563	5.3037E-03
YBR084W	Mis1p	-0.3558	1.1816E-03
YDR099W	Bmh2p	-0.3533	4.6161E-04
YPR128C	Ant1p	-0.3525	9.6078E-03
YMR300C	Ade4p	-0.3511	1.4854E-04
YDL161W	Ent1p	-0.3509	9.4179E-04
YHL007C	Ste20p	-0.3507	1.2379E-03
YIR037W	Hyr1p	-0.3445	1.0498E-04
YKR065C	Pam17p	-0.3437	8.3424E-04
YPR029C	Apl4p	-0.3419	2.6197E-03
YGR208W	Ser2p	-0.3412	2.8467E-03
YOR341W	Rpa190p	-0.3411	6.6265E-04
YGR218W	Crm1p	-0.3399	1.2034E-04
YNL110C	Nop15p	-0.3392	8.3424E-04
YKL054C	Def1p	-0.3363	4.0249E-04
YOR212W	Ste4p	-0.3363	7.4578E-03
YGL099W	Lsg1p	-0.3352	2.0772E-04
YER025W	Gcd11p	-0.3325	4.5200E-05
YMR296C	Lcb1p	-0.3316	4.4909E-03
YBR112C	Cyc8p	-0.3315	4.3462E-03
YOR217W	Rfc1p	-0.3295	3.2409E-03
YPR073C	Ltp1p	-0.3289	5.3180E-03
YMR243C	Zrc1p	-0.3284	1.0498E-04
YIL053W	Gpp1p	-0.3273	3.5019E-03

YDR238C	Sec26p	-0.3265	1.8544E-04
YJL158C	Cis3p	-0.3254	8.5650E-03
YOR326W	Myo2p	-0.3237	9.5816E-03
YBR205W	Ktr3p	-0.3237	6.8100E-05
YDL040C	Nat1p	-0.3237	7.6805E-04
YPL093W	Nog1p	-0.3234	1.6086E-04
YBR061C	Trm7p	-0.3232	2.3278E-03
YKL164C	Pir1p	-0.3230	1.8559E-03
YER182W	Fmp10p	-0.3218	5.1243E-04
YDL174C	Dld1p	-0.3215	1.4546E-04
YDR207C	Ume6p	-0.3199	6.3306E-03
YOR091W	Tma46p	-0.3191	1.4189E-04
YKL013C	Arc19p	-0.3183	5.0876E-03
YER034W		-0.3179	6.5219E-03
YBL091C	Map2p	-0.3167	9.2092E-04
YOR133W	Eft1p	-0.3164	6.6900E-05
YDR385W	Eft2p	-0.3164	6.6900E-05
YMR183C	Sso2p	-0.3127	3.9359E-04
YPR133C	Spn1p	-0.3116	1.8544E-04
YBL024W	Ncl1p	-0.3083	2.2104E-04
YPL063W	Tim50p	-0.3066	4.3152E-04
YJR132W	Nmd5p	-0.3030	4.8323E-03
YDR341C		-0.3026	5.7154E-04
YCR077C	Pat1p	-0.3023	1.7669E-03
YOL059W	Gpd2p	-0.3004	9.0661E-04
YNL247W		-0.2998	1.3045E-04
YMR109W	Myo5p	-0.2991	4.5288E-03
YPR165W	Rho1p	-0.2978	7.6513E-04
YOL109W	Zeo1p	-0.2967	2.8526E-03
YKL215C	Oxp1p	-0.2965	4.6514E-03
YOL111C	Mdy2p	-0.2959	2.9362E-03
YKL082C	Rrp14p	-0.2950	4.5971E-04
YLR179C		-0.2898	1.1486E-03
YIL070C	Mam33p	-0.2893	2.2692E-04
YDL031W	Dbp10p	-0.2890	5.0046E-03
YNL007C	Sis1p	-0.2871	1.2832E-04
YGR078C	Pac10p	-0.2867	1.4334E-03
YMR083W	Adh3p	-0.2849	7.1496E-04
YML106W	Ura5p	-0.2832	1.6807E-04

YCR084C	Tup1p	-0.2827	1.3632E-04
YLR359W	Ade13p	-0.2821	1.9379E-04
YER081W	Ser3p	-0.2791	1.2867E-03
YDR381W	Yra1p	-0.2791	1.6487E-04
YHR183W	Gnd1p	-0.2781	1.4189E-04
YPL243W	Srp68p	-0.2781	5.3037E-03
YBR088C	Pol30p	-0.2781	1.9890E-04
YGR193C	Pdx1p	-0.2772	1.8224E-04
YPL145C	Kes1p	-0.2734	2.2692E-04
YNL208W		-0.2718	1.4666E-04
YEL047C	Frd1p	-0.2716	3.1803E-04
YPL129W	Taf14p	-0.2697	3.5754E-04
YBL007C	Sla1p	-0.2696	1.7669E-03
YPR010C	Rpa135p	-0.2688	5.3037E-03
YOL077C	Brx1p	-0.2670	7.4158E-03
YKL212W	Sac1p	-0.2669	7.0346E-04
YJR014W	Tma22p	-0.2665	1.8386E-03
YCR009C	Rvs161p	-0.2656	3.3111E-04
YNL016W	Pub1p	-0.2638	1.6522E-03
YIL125W	Kgd1p	-0.2629	3.7632E-04
YKL172W	Ebp2p	-0.2612	5.7737E-03
YHR066W	Ssf1p	-0.2607	2.4626E-03
YDL122W	Ubp1p	-0.2596	4.8262E-03
YNR051C	Bre5p	-0.2593	5.8027E-03
YHL033C	Rpl8Ap	-0.2568	1.8051E-04
YLR301W	Hri1p	-0.2564	6.4538E-04
YNR035C	Arc35p	-0.2564	1.1903E-03
YDR388W	Rvs167p	-0.2552	1.4686E-03
YKL032C	Ixr1p	-0.2548	1.9815E-03
YOR283W		-0.2531	4.0973E-03
YGR204W	Ade3p	-0.2529	1.5675E-04
YOL056W	Gpm3p	-0.2515	8.5644E-03
YPR183W	Dpm1p	-0.2503	2.7806E-04
YBR196C	Pgi1p	-0.2496	5.6077E-04
YIL083C	Cab2p	-0.2491	1.8329E-03
YBR035C	Pdx3p	-0.2488	1.5487E-04
YHR104W	Gre3p	-0.2486	1.9311E-03
YER036C	Arb1p	-0.2476	1.2481E-04
YJR059W	Ptk2p	-0.2468	2.3731E-03

YDR060W	Mak21p	-0.2462	9.6444E-04
YMR273C	Zds1p	-0.2450	3.2409E-03
YGR155W	Cys4p	-0.2440	4.0251E-04
YPL106C	Sse1p	-0.2423	8.8301E-04
YOL051W	Gal11p	-0.2407	7.6513E-04
YPR181C	Sec23p	-0.2403	2.1263E-04
YKR001C	Vps1p	-0.2385	1.5404E-03
YPL212C	Pus1p	-0.2379	2.5662E-03
YPL235W	Rvb2p	-0.2365	1.6086E-04
YER148W	Spt15p	-0.2362	3.3613E-03
YJL111W	Cct7p	-0.2356	3.5136E-04
YCL028W	Rnq1p	-0.2352	5.3203E-03
YNL121C	Tom70p	-0.2352	9.7496E-04
YNR050C	Lys9p	-0.2346	1.5462E-03
YDR116C	Mrpl1p	-0.2335	2.0352E-03
YOR164C	Get4p	-0.2328	2.1229E-04
YGL055W	Ole1p	-0.2327	2.7141E-03
YLR372W	Sur4p	-0.2299	5.1544E-03
YJL081C	Arp4p	-0.2296	4.4440E-03
YER107C	Gle2p	-0.2294	9.0478E-03
YPR040W	Tip41p	-0.2294	9.8699E-03
YGR119C	Nup57p	-0.2282	2.2314E-04
YJL186W	Mnn5p	-0.2274	1.6505E-03
YDR101C	Arx1p	-0.2242	8.3424E-04
YGR162W	Tif4631p	-0.2242	2.2692E-04
YAL060W	Bdh1p	-0.2239	8.5473E-03
YMR315W		-0.2225	6.3982E-03
YIL021W	Rpb3p	-0.2216	9.2692E-03
YJR060W	Cbf1p	-0.2211	8.3424E-04
YGR116W	Spt6p	-0.2195	2.9541E-04
YML071C	Cog8p	-0.2186	6.6362E-03
YJL026W	Rnr2p	-0.2184	8.9813E-04
YOR239W	Abp140p	-0.2180	1.8415E-03
YFL039C	Act1p	-0.2172	4.4073E-04
YMR309C	Nip1p	-0.2163	2.0485E-04
YGL087C	Mms2p	-0.2159	5.9226E-03
YDL135C	Rdi1p	-0.2158	2.0485E-04
YOR361C	Prt1p	-0.2153	5.6579E-04
YLR448W	Rpl6Bp	-0.2151	7.1334E-04

YBR154C	Rpb5p	-0.2145	3.1872E-03
YAL038W	Cdc19p	-0.2134	1.0641E-03
YML012W	Erv25p	-0.2131	7.0584E-03
YPR148C		-0.2127	6.9316E-03
YLL011W	Sof1p	-0.2122	2.1583E-03
YER125W	Rsp5p	-0.2106	6.5700E-03
YER048C	Caj1p	-0.2105	5.8324E-04
YBR121C	Grs1p	-0.2104	2.6352E-03
YER087C-B	Sbh1p	-0.2104	6.6362E-03
YLR109W	Ahp1p	-0.2097	5.6954E-04
YDR486C	Vps60p	-0.2087	3.6846E-03
YJL055W		-0.2068	3.4326E-03
YNL153C	Gim3p	-0.2040	3.5529E-04
YLR378C	Sec61p	-0.2023	3.9541E-03
YDR189W	Sly1p	-0.2012	5.6711E-03
YJR105W	Ado1p	-0.2006	1.3954E-03
YJL101C	Gsh1p	-0.1993	5.9044E-04
YOL041C	Nop12p	-0.1971	8.6969E-03
YLR044C	Pdc1p	-0.1968	7.4561E-03
YOR310C	Nop58p	-0.1931	1.4537E-03
YDR025W	Rps11Ap	-0.1919	1.3458E-03
YBR048W	Rps11Bp	-0.1919	1.3458E-03
YLR216C	Cpr6p	-0.1913	1.3778E-03
YBL050W	Sec17p	-0.1905	1.5462E-03
YDR429C	Tif35p	-0.1901	1.7707E-03
YNL292W	Pus4p	-0.1898	5.0631E-03
YKL024C	Ura6p	-0.1888	2.3312E-03
YDL053C	Pbp4p	-0.1879	4.5148E-03
YBL026W	Lsm2p	-0.1877	6.1743E-03
YPR154W	Pin3p	-0.1869	7.7452E-03
YDL185W	Vma1p	-0.1847	8.3654E-03
YPR035W	Gln1p	-0.1841	1.1486E-03
YOL094C	Rfc4p	-0.1825	5.8008E-03
YLR118C		-0.1791	4.2546E-03
YER133W	Glc7p	-0.1786	4.3313E-03
YKL009W	Mrt4p	-0.1776	1.5938E-03
YML079W		-0.1776	6.3093E-03
YMR311C	Glc8p	-0.1760	3.7046E-03
YDL103C	Qri1p	-0.1737	3.2409E-03

YAL007C	Erp2p	-0.1709	5.9794E-03
YGL244W	Rtf1p	-0.1681	3.5927E-03
YNL209W	Ssb2p	-0.1677	2.9362E-03
YKL006C-A	Sft1p	-0.1661	3.2172E-03
YNL307C	Mck1p	-0.1655	3.5310E-03
YHL034C	Sbp1p	-0.1651	9.2692E-03
YNL067W	Rpl9Bp	-0.1646	6.8005E-03
YER009W	Ntf2p	-0.1641	5.3037E-03
YMR153W	Nup53p	-0.1634	4.0973E-03
YGL207W	Spt16p	-0.1606	3.2294E-03
YPL160W	Cdc60p	-0.1598	4.3462E-03
YIL075C	Rpn2p	-0.1589	5.3203E-03
YER089C	Ptc2p	-0.1581	4.5022E-03
YER136W	Gdi1p	-0.1578	2.1785E-03
YJR069C	Ham1p	-0.1551	1.7539E-03
YHR121W	Lsm12p	-0.1519	4.3084E-03
YMR161W	Hlj1p	-0.1516	4.7747E-03
YBR106W	Pho88p	-0.1511	3.0487E-03
YIL093C	Rsm25p	-0.1505	3.6846E-03
YGL008C	Pma1p	-0.1499	9.8881E-03
YDR427W	Rpn9p	-0.1491	1.6203E-03
YDL055C	Psa1p	-0.1478	8.1899E-03
YML126C	Erg13p	-0.1471	1.7614E-03
YKL193C	Sds22p	-0.1470	6.9725E-03
YMR146C	Tif34p	-0.1424	2.1525E-03
YDL205C	Hem3p	-0.1358	8.1899E-03
YDR510W	Smt3p	-0.1317	8.4459E-03
YAR002C-A	Erp1p	-0.1305	6.3421E-03
YJL008C	Cct8p	-0.1261	5.7223E-03
YGR285C	Zuo1p	-0.1253	8.0464E-03
YCR002C	Cdc10p	-0.1235	5.3203E-03
YPL090C	Rps6Ap	-0.1227	5.2956E-03
YBR181C	Rps6Bp	-0.1227	5.2956E-03
YPL178W	Cbc2p	-0.1179	6.2778E-03
YOR063W	Rpl3p	-0.1127	4.8992E-03
YPR074C	Tkl1p	-0.1002	8.1899E-03

Gene ontology results for up regulated proteins in elp3/wt at 1%FDR (followed by list of proteins)

funspec: P-value cutoff=0.01 with Bonferroni correction				
GO Molecular Function (1646 categories)				
Category	**p-value**	**In Category from Cluster**	**k**	**f**
threonine-type endopeptidase activity [GO:0004298]	1.03E-11	PRE1 PUP3 SCL1 PRE9 PUP2 PRE3 PRE8 PRE6 PUP1 PRE10 PRE2	11	14
catalytic activity [GO:0003824]	1.50E-10	CYS3 COR1 ETR1 PTC4 HIS7 NFS1 THR4 PTC1 HNT1 LYS21 LYS20 ARO3 ARO10 UBA2 UTR4 FMP52 ARG5,6 YER152C BNA6 ERG26 TRP5 SCW11 ARO8 LSC2 CPD1 SCW4 BGL2 YHI9 PTC7 UBA4 ARO9 OYE2 ERG9 IMD2 HIS6 BNA3 ILV3 CPA2 MAE1 UBA1 URA1 YKR070W MTD1 MEU1 AAT2 ALT1 PNP1 CTS1 EXG1 ARG7 YMR226C FAA4 SCW10 ADH6 IDH1 LEU4 RIO2 ZWF1 LSC1 ERG10 FCY1 SPE3 QCR2	63	455
endopeptidase activity [GO:0004175]	1.59E-10	PRE1 PUP3 SCL1 PRE9 PUP2 RPN1 PRE3 PRE8 PRE6 PUP1 PRE10 PRE2	12	20
oxidoreductase activity [GO:0016491]	7.39E-07	COX2 PRX1 ETR1 MXR2 IDP1 YFH1 YPR1 MXR1 ARG5,6 ALD5 LPD1 ERG26 ERG1 AIM17 SOD2 OYE2 ERG9 IMD2 BNA1 SOD1 MAE1 URA1 CCP1 MTD1 TSA1 PGA3 ERO1 NDE1 YMR226C ADH6 IDH1 ZWF1 CAT5 PRO2 CIR2 ALD4 GLR1 YPR127W	38	272
metal ion binding [GO:0046872]	1.08E-06	COX2 COR1 PTC4 MXR2 PRD1 YCR087C-A PTC1 RPS29B IDP1 GCS1 LYS4 ARO10 UBA2 RIB3 UTR4 PCM1 PMI40 PRS2 MDJ1 OTU1 YFR006W DUG1 COX4 SPT4 ZPR1 PRS3 AIM17 SOD2 COX6 PTC7 UBA4 IMD2 AGE2 ACO2 ILV3 BNA1 SOD1 CPA2 MAE1 MSN4 CCP1 DRE2 SAM1 MAP1 URA4 GLO1 APT1 IRC21 SCJ1 ADH6 NCE103 IDH1 YDJ1 DMA2 ADE12 PRS5 MET22 CYT1 CAT5 CIR2 ERG10 GRX5 CAR1 IDI1 ISU1 SAM4 FCY1 SUA7 QCR2	69	647
proton-transporting ATPase activity, rotational mechanism [GO:0046961]	4.02E-06	ATP3 ATP5 TIM11 VMA7 ATP2 ATP7 ATP14 VMA6 ATP4 ATP15	10	29

GO Biological Process (2062 categories)				
Category	**p-value**	**In Category from Cluster**	**k**	**f**
proteasomal ubiquitin-independent protein catabolic process [GO:0010499]	1.03E-11	PRE1 PUP3 SCL1 PRE9 PUP2 PRE3 PRE8 PRE6 PUP1 PRE10 PRE2	11	14
cellular amino acid biosynthetic process [GO:0008652]	5.88E-11	CYS3 HIS7 ILV6 THR4 LYS21 LYS20 ARO3 LYS4 TRP4 UTR4 HOM3 HIS1 ARG5,6 TRP2 TRP5 ARO2 ASN2 HIS6 ILV3 CPA2 ARG7 LEU4 ARG1 PRO2 SAM4 ASN1	26	98
metabolic process [GO:0008152]	6.80E-11	ETR1 HIS7 ILV6 NFS1 THR4 HNT1 LYS21 LYS20 DAS2 ARO3 PAA1 LYS4 TRP4 UBA2 UTR4 FMP52 HOM3 ARG5,6 ALD5 DUG1 BNA6 ERG26 TRP5 SCW11 ERG1 LSC2 SCW4 BGL2 PTC7 UBA4 OYE2 ERG9 IMD2 HIS6 ACO2 ILV3 BNA1 CPA2 MAE1 UBA1 URA1 YKR070W MTD1 CTS1 EXG1 UNG1 ARG7 YMR226C FAA4 SCW10 ADH6 IDH1 LEU4 ZWF1 DSE4 LSC1 PRO2 ALD4 ERG10 POS5 FCY1	61	425
proteolysis involved in cellular protein catabolic process [GO:0051603]	7.23E-10	PRE1 PUP3 SCL1 PRE9 PUP2 PRE3 PRE8 PRE6 PUP1 PRE10 PRE2	11	18
proteasomal ubiquitin-dependent protein catabolic process [GO:0043161]	1.16E-09	SHP1 SEM1 PRE1 PUP3 RPN11 SCL1 PRE9 PUP2 PRE3 PRE8 PRE6 PUP1 PRE10 PRE2	14	32
ubiquitin-dependent protein catabolic process [GO:0006511]	1.23E-08	RPT2 RPN6 CDC48 CDC53 SEM1 RPT3 DDI1 UBP6 RPN12 SCL1 PRE9 PUP2 RPN1 RPN10 RPT1 PRE8 PRE6 RPT5 PRE10	19	69
protein folding [GO:0006457]	1.68E-07	PDI1 TAH1 CCT4 TCP1 FPR2 MDJ1 PHB1 PHB2 ESS1 SSC1 SBA1 SSA2 TSA1 CPR3 ERO1 HSC82 SCJ1 YDJ1 HSP10 STI1 HSP82	21	96
oxidation-reduction process [GO:0055114]	7.39E-07	COX2 PRX1 ETR1 MXR2 IDP1 YFH1 YPR1 MXR1 ARG5,6 ALD5 LPD1 ERG26 ERG1 AIM17 SOD2 OYE2 ERG9 IMD2 BNA1 SOD1 MAE1 URA1 CCP1 MTD1 TSA1 PGA3 ERO1 NDE1 YMR226C ADH6 IDH1 ZWF1 CAT5 PRO2 CIR2 ALD4 GLR1 YPR127W	38	272
biosynthetic process [GO:0009058]	2.33E-06	ARO3 GCD6 TRP2 YER152C ARO8 CAB4 YHI9 ARO9 ERG9 BNA3 AAT2 ALT1	12	40
ATP synthesis coupled proton transport [GO:0015986]	2.44E-06	ATP3 ATP5 TIM11 ATP2 ATP7 ATP14 ATP4 ATP15	8	17

MIPS Functional Classification (459 categories)				
Category	**p-value**	**In Category from Cluster**	**k**	**f**
protein processing (proteolytic) [14.07.11]	3.79E-11	RPT2 RPT3 PRE1 PUP3 RPN11 RPN12 SCL1 PRE9 PUP2 RPN1 RPN10 PRE3 RPT1 MAP1 PRE8 PRE6 RPT5 PUP1 RPT4 PRE10 PRE2	21	63
proteasomal degradation (ubiquitin/proteasomal pathway) [14.13.01.01]	2.88E-10	SHP1 RPT2 RPN6 CDC48 CDC53 SEM1 RPT3 PRE1 PUP3 DDI1 OTU1 RPN11 RPN12 SCL1 PRE9 PUP2 RPN1 RPN10 PRE3 RPT1 UBA1 SEC13 PRE8 PRE6 RPT5 PUP1 RPT4 PRE10 PRE2	29	128
electron transport [20.01.15]	1.87E-09	COX2 ATP3 ATP5 TIM11 ARO2 COX4 VMA7 COX6 OYE2 ATP2 ATP7 URA1 CCP1 COX12 ATP14 VMA6 ERO1 NDE1 CYT1 GRX5 ATP4 ATP15	22	83
electron transport and membrane-associated energy conservation [02.11]	2.61E-08	COX2 COR1 ATP3 ATP5 TIM11 QCR7 ALD5 COX4 COX6 ATP2 ATP7 COX12 ATP14 NDE1 ATP4 ATP15 QCR2	17	58
protein folding and stabilization [14.01]	4.45E-07	PDI1 TAH1 CCT4 CDC37 TCP1 HSP78 FPR2 MDJ1 ESS1 SSC1 SBA1 SSA2 CPR3 ERO1 HSC82 SCJ1 YDJ1 HSP10 STI1 HSP82	20	93
energy generation (e.g. ATP synthase) [02.45.15]	1.48E-06	ATP3 ATP5 TIM11 ATP2 ATP7 ATP14 CYT1 ATP4 ATP15	9	21
biosynthesis of tryptophan [01.01.09.06.01]	1.65E-05	TRP4 TRP2 PRS2 TRP5 PRS3 PRS5	6	11
MIPS Phenotypes (142 categories)				
Category	**p-value**	**In Category from Cluster**	**k**	**f**
Respiratory deficiency [42.25.20]	3.93E-06	COX2 PIM1 PET9 ATP3 PTC1 YFH1 ATP5 TIM11 QCR7 MDJ1 LSC2 MMF1 ACP1 COX12 ATP14 CPR3 SEC65 COX14 NDE1 IDH1 POR1 CIT1 CAT5 ATP4 ATP15 SRP54 QCR2	27	173

List of all proteins up regulated in elp3/wt at 1%FDR

Systematic Name	Standard Name	logFC	adj.P.Val
YHR137W	Aro9p	2.0511	2.4100E-06
YDR380W	Aro10p	1.9244	1.5000E-05
YJR109C	Cpa2p	1.3913	2.9300E-05
YHR216W	Imd2p	1.3522	4.0200E-05
YKR042W	Uth1p	1.2639	3.6846E-03
YJL200C	Aco2p	1.2349	1.3570E-04
YOR374W	Ald4p	1.2211	2.2202E-04
YLR286C	Cts1p	1.2073	1.2175E-04
YER073W	Ald5p	1.1444	2.0245E-04
YGL117W		1.1144	4.9380E-04
YMR062C	Arg7p	1.0767	3.6700E-05
YOR125C	Cat5p	1.0233	5.1400E-05
YGL028C	Scw11p	1.0153	2.0093E-03
YML004C	Glo1p	0.9966	1.2781E-04
YKL062W	Msn4p	0.9760	5.4045E-04
YDR234W	Lys4p	0.9716	3.3000E-05
YDR258C	Hsp78p	0.9190	4.9900E-06
YDL131W	Lys21p	0.8956	3.3000E-05
YDL182W	Lys20p	0.8824	4.2300E-05
YMR305C	Scw10p	0.8535	4.4073E-04
YHR029C	Yhi9p	0.8346	4.9900E-06
YLR300W	Exg1p	0.8057	4.9900E-06
YPR145W	Asn1p	0.8024	7.3299E-04
YMR318C	Adh6p	0.7958	2.1174E-04
YBR256C	Rib5p	0.7940	4.3152E-04
YJL191W	Rps14Bp	0.7932	3.3791E-04
YKR076W	Ecm4p	0.7821	3.0388E-03
YGR247W	Cpd1p	0.7468	4.9367E-04
YER152C		0.7467	2.1602E-03
YOR176W	Hem15p	0.7438	1.0194E-03
YOR222W	Odc2p	0.7318	3.6700E-05
YKL216W	Ura1p	0.7258	1.7200E-05
YER052C	Hom3p	0.7103	1.6708E-04
YEL060C	Prb1p	0.7086	6.7300E-05
YCR087C-A		0.7043	2.0178E-04
YNR046W	Trm112p	0.7024	3.6700E-05
YPL273W	Sam4p	0.7009	5.0513E-03
YOL143C	Rib4p	0.6981	7.4900E-05

YBL045C	Cor1p	0.6977	7.6300E-06
YHL021C	Aim17p	0.6546	3.0600E-05
YER055C	His1p	0.6495	3.1000E-05
YPL135W	Isu1p	0.6432	5.1748E-04
YPL245W		0.6393	1.2850E-04
YAL044C	Gcv3p	0.6371	1.6952E-03
YKR048C	Nap1p	0.6221	1.1400E-05
YER004W	Fmp52p	0.6169	3.1000E-05
YNL149C	Pga2p	0.6158	7.0000E-06
YOR157C	Pup1p	0.6148	5.9044E-04
YDR365C	Esf1p	0.6146	2.8100E-05
YPR127W		0.6096	8.7758E-03
YLR055C	Spt8p	0.6091	1.4046E-04
YNL104C	Leu4p	0.6087	4.3152E-04
YKR066C	Ccp1p	0.6067	1.7200E-05
YBR101C	Fes1p	0.5967	5.0400E-05
YER042W	Mxr1p	0.5963	1.1153E-04
YKL126W	Ypk1p	0.5934	4.2100E-05
YER090W	Trp2p	0.5928	1.1600E-05
YDL173W	Par32p	0.5884	8.7652E-04
YLR038C	Cox12p	0.5810	6.3900E-05
YJR103W	Ura8p	0.5803	3.1085E-03
YOR362C	Pre10p	0.5800	2.7600E-05
YBL058W	Shp1p	0.5781	5.7700E-05
YFL028C	Caf16p	0.5761	1.2542E-04
YFR006W		0.5704	8.3500E-05
YGL026C	Trp5p	0.5694	7.0000E-06
YDR019C	Gcv1p	0.5658	8.7300E-05
YMR178W		0.5605	3.2900E-05
YEL056W	Hat2p	0.5561	5.1243E-04
YOR259C	Rpt4p	0.5486	2.0900E-05
YPL170W	Dap1p	0.5463	1.3632E-04
YFL016C	Mdj1p	0.5462	2.3800E-05
YDL125C	Hnt1p	0.5429	2.9344E-04
YOR367W	Scp1p	0.5428	1.2379E-03
YDL006W	Ptc1p	0.5425	2.9512E-04
YHL011C	Prs3p	0.5386	2.3300E-05
YGR063C	Spt4p	0.5375	3.1000E-05
YDR354W	Trp4p	0.5354	1.7381E-04

YCR060W	Tah1p	0.5352	1.5700E-05
YDL198C	Ggc1p	0.5336	7.6513E-04
YDR487C	Rib3p	0.5319	1.4854E-04
YDR394W	Rpt3p	0.5258	1.0500E-05
YMR226C		0.5243	3.6700E-05
YHL031C	Gos1p	0.5241	3.3200E-05
YDR390C	Uba2p	0.5210	5.2300E-05
YOR251C	Tum1p	0.5183	1.3000E-05
YJL001W	Pre3p	0.5166	1.6708E-04
YNL036W	Nce103p	0.5147	1.4189E-04
YLR257W		0.5138	3.2400E-05
YPR086W	Sua7p	0.5109	4.8500E-05
YKL040C	Nfu1p	0.5109	2.7600E-05
Q0250	Cox2p	0.5098	1.9205E-03
YER094C	Pup3p	0.5097	9.4400E-05
YHR051W	Cox6p	0.5060	1.1100E-05
YEL032W	Mcm3p	0.5052	1.1968E-04
YPR103W	Pre2p	0.5043	1.6708E-04
YBL022C	Pim1p	0.5039	5.6800E-05
YPL117C	Idi1p	0.4994	1.4298E-04
YKL191W	Dph2p	0.4963	2.9344E-04
YDR035W	Aro3p	0.4960	1.2040E-04
YER143W	Ddi1p	0.4947	5.0575E-04
YGR244C	Lsc2p	0.4910	5.8100E-05
YGR038W	Orm1p	0.4907	3.6700E-05
YNR067C	Dse4p	0.4843	5.5401E-04
YBR125C	Ptc4p	0.4828	4.9380E-04
YFL044C	Otu1p	0.4808	1.7381E-04
YDR531W	Cab1p	0.4806	9.5557E-04
YHR076W	Ptc7p	0.4804	1.2542E-04
YDR129C	Sac6p	0.4784	2.4669E-04
YDL120W	Yfh1p	0.4777	3.3006E-03
YPR191W	Qcr2p	0.4776	3.9600E-05
YOR356W	Cir2p	0.4736	1.3325E-04
YCL009C	Ilv6p	0.4662	5.0400E-05
YJR016C	Ilv3p	0.4661	8.5451E-03
YHR190W	Erg9p	0.4646	1.7300E-05
YER099C	Prs2p	0.4616	3.1085E-03
YIL008W	Urm1p	0.4592	2.6302E-04

YOR209C	Npt1p	0.4588	2.8300E-05
YPL059W	Grx5p	0.4570	2.7600E-05
YBR177C	Eht1p	0.4563	1.0988E-04
YDL007W	Rpt2p	0.4562	1.1570E-04
YHR027C	Rpn1p	0.4534	1.6700E-05
YML101C	Cue4p	0.4479	3.1234E-04
YDR056C		0.4478	1.6408E-03
YPR016C	Tif6p	0.4473	1.0453E-04
YOR370C	Mrs6p	0.4466	5.4500E-05
YKL029C	Mae1p	0.4438	5.6077E-04
YIL002W-A		0.4408	8.8700E-05
YKL007W	Cap1p	0.4388	3.6700E-05
YMR236W	Taf9p	0.4380	6.0011E-04
YKL145W	Rpt1p	0.4341	1.4189E-04
YLR025W	Snf7p	0.4331	4.5200E-05
YMR276W	Dsk2p	0.4321	2.8300E-05
YNL035C		0.4318	2.9332E-03
YER156C		0.4302	1.1204E-04
YNR001C	Cit1p	0.4299	3.2400E-05
YML130C	Ero1p	0.4297	2.9541E-04
YKL096W	Cwp1p	0.4282	1.1283E-03
YKL192C	Acp1p	0.4249	1.6905E-03
YOR229W	Wtm2p	0.4235	3.1872E-03
YBR026C	Etr1p	0.4232	1.0453E-04
YHR111W	Uba4p	0.4227	1.2839E-03
YFR049W	Ymr31p	0.4204	2.0627E-04
YGL187C	Cox4p	0.4191	1.6962E-04
YDR211W	Gcd6p	0.4157	1.0397E-04
YOR323C	Pro2p	0.4152	9.4400E-05
YPR034W	Arp7p	0.4149	3.6700E-05
YML021C	Ung1p	0.4140	1.5641E-04
YPR069C	Spe3p	0.4139	2.4900E-05
YMR214W	Scj1p	0.4118	3.6020E-03
YJR125C	Ent3p	0.4103	5.2165E-04
YNL115C		0.4099	1.1129E-03
YLR209C	Pnp1p	0.4093	3.7129E-03
YJR104C	Sod1p	0.4080	6.8100E-05
YEL058W	Pcm1p	0.4079	2.7600E-05
YOR117W	Rpt5p	0.4018	2.2000E-05

YOR115C	Trs33p	0.4014	8.8700E-05
YIL041W	Gvp36p	0.4006	2.7600E-05
YLR225C		0.4002	1.0444E-03
YNL241C	Zwf1p	0.3984	2.7600E-05
YDR517W	Grh1p	0.3983	3.7600E-05
YGR211W	Zpr1p	0.3973	3.1872E-03
YHL039W	Efm1p	0.3966	1.4224E-03
YCL043C	Pdi1p	0.3966	1.1702E-04
YPL239W	Yar1p	0.3948	3.7447E-04
YDR020C	Das2p	0.3948	5.1028E-03
YLR370C	Arc18p	0.3931	7.8561E-04
YAL042W	Erv46p	0.3931	1.6994E-04
YHR012W	Vps29p	0.3915	1.1310E-03
YNL220W	Ade12p	0.3912	1.3395E-04
YGR132C	Phb1p	0.3907	3.4000E-05
YFR044C	Dug1p	0.3887	4.8800E-05
YLR212C	Tub4p	0.3881	4.0071E-03
YEL038W	Utr4p	0.3870	1.9890E-04
YLR195C	Nmt1p	0.3825	1.6983E-04
YCR059C	Yih1p	0.3817	1.2379E-03
YER183C	Fau1p	0.3807	4.7845E-03
YKR080W	Mtd1p	0.3723	1.8463E-04
YKL210W	Uba1p	0.3698	9.0900E-05
YGL106W	Mlc1p	0.3690	2.8000E-05
YGR124W	Asn2p	0.3672	2.7600E-05
YCL011C	Gbp2p	0.3671	2.9400E-05
YGL221C	Nif3p	0.3665	9.4134E-04
YLL001W	Dnm1p	0.3663	1.8544E-04
YAR014C	Bud14p	0.3630	5.2131E-03
YCL057W	Prd1p	0.3630	5.6579E-04
YNL207W	Rio2p	0.3602	7.1334E-04
YOL058W	Arg1p	0.3601	6.0709E-03
YBL064C	Prx1p	0.3585	4.3228E-03
YGL202W	Aro8p	0.3555	6.4300E-05
YOR243C	Pus7p	0.3548	1.4662E-04
YML058W	Sml1p	0.3525	2.3724E-03
YGR001C		0.3521	2.1143E-03
YBR122C	Mrpl36p	0.3494	6.8371E-03
YFR004W	Rpn11p	0.3485	1.9859E-04

YGL242C		0.3478	7.7386E-04
YGR282C	Bgl2p	0.3476	1.3679E-04
YDR322C-A	Tim11p	0.3468	2.2683E-03
YIL044C	Age2p	0.3457	3.7615E-04
YPL012W	Rrp12p	0.3454	5.9226E-03
YDR363W-A	Sem1p	0.3453	1.7263E-03
YCR073W-A	Sol2p	0.3438	7.8257E-04
YLR208W	Sec13p	0.3437	8.8700E-05
YFR047C	Bna6p	0.3435	7.9220E-03
YGR231C	Phb2p	0.3426	6.8100E-05
YDR529C	Qcr7p	0.3423	1.0397E-04
YPR062W	Fcy1p	0.3408	6.9042E-04
YNR032C-A	Hub1p	0.3402	2.1473E-03
YIL020C	His6p	0.3395	1.9572E-03
YIL051C	Mmf1p	0.3387	9.0464E-04
YML092C	Pre8p	0.3385	3.9300E-05
YGL112C	Taf6p	0.3380	1.1740E-03
YIL034C	Cap2p	0.3376	3.0833E-04
YGR086C	Pil1p	0.3371	5.0513E-03
YKL130C	She2p	0.3364	1.5977E-03
YOL052C	Spe2p	0.3349	7.4067E-03
YMR073C	Irc21p	0.3347	1.7445E-03
YOL070C	Nba1p	0.3309	1.5977E-03
YBR230C	Om14p	0.3303	9.6444E-04
YBR162C	Tos1p	0.3287	2.3613E-03
YBL047C	Ede1p	0.3285	2.6424E-03
YAL025C	Mak16p	0.3274	3.6303E-03
YDR168W	Cdc37p	0.3265	1.2850E-04
YDR361C	Bcp1p	0.3258	3.8181E-03
YHR008C	Sod2p	0.3253	3.4157E-04
YHR179W	Oye2p	0.3246	8.3810E-03
YML129C	Cox14p	0.3235	1.0232E-04
YOL149W	Dcp1p	0.3226	3.8441E-04
YNR029C		0.3223	7.1496E-04
YJL123C	Mtc1p	0.3217	1.0067E-04
YEL036C	Anp1p	0.3213	8.0845E-04
YLR244C	Map1p	0.3197	5.0575E-04
YDR368W	Ypr1p	0.3196	2.2202E-04
YKR071C	Dre2p	0.3186	2.9018E-03

YNL055C	Por1p	0.3172	5.2244E-04
YKR070W		0.3152	8.9813E-04
YGL148W	Aro2p	0.3149	1.0758E-03
YLR017W	Meu1p	0.3143	1.0665E-04
YNR053C	Nog2p	0.3139	3.4157E-04
YPL188W	Pos5p	0.3132	3.3370E-03
YPL111W	Car1p	0.3112	1.0067E-04
YPR088C	Srp54p	0.3112	1.2728E-04
YOR007C	Sgt2p	0.3075	5.6077E-04
YPL169C	Mex67p	0.3070	8.6795E-03
YMR145C	Nde1p	0.3057	7.4647E-03
YMR002W	Mic17p	0.3057	7.5263E-03
YDR167W	Taf10p	0.3055	1.2040E-04
YNL287W	Sec21p	0.3042	9.4341E-04
YOR197W	Mca1p	0.3037	5.2803E-04
YGR277C	Cab4p	0.3031	6.5044E-03
YPL211W	Nip7p	0.3027	2.9560E-03
YPR041W	Tif5p	0.3025	2.2381E-04
YOR298C-A	Mbf1p	0.3011	3.7725E-04
YPL240C	Hsp82p	0.2995	3.7615E-04
YJR144W	Mgm101p	0.2985	2.6278E-03
YKR018C		0.2962	6.9327E-04
YKL142W	Mrp8p	0.2961	6.8317E-03
YDL226C	Gcs1p	0.2935	1.0453E-04
YOR286W	Rdl2p	0.2920	4.9380E-04
YHR200W	Rpn10p	0.2917	9.8100E-05
YGR279C	Scw4p	0.2910	2.5808E-03
YDL066W	Idp1p	0.2896	1.0728E-04
YFL018C	Lpd1p	0.2872	2.5882E-04
YGL100W	Seh1p	0.2837	1.0623E-03
YGR020C	Vma7p	0.2826	2.2202E-04
YJL069C	Utp18p	0.2821	6.0551E-04
YGR253C	Pup2p	0.2815	3.9397E-04
YML105C	Sec65p	0.2811	1.0952E-03
YDL192W	Arf1p	0.2804	2.9541E-04
YKL117W	Sba1p	0.2799	1.0498E-04
YER012W	Pre1p	0.2793	7.6805E-04
YJR085C		0.2793	3.2623E-03
YAL012W	Cys3p	0.2762	4.9031E-04

YHR199C	Aim46p	0.2758	1.0308E-03
YOR224C	Rpb8p	0.2726	1.5490E-04
YKL094W	Yju3p	0.2715	9.2092E-04
YLR420W	Ura4p	0.2715	3.2294E-03
YJR065C	Arp3p	0.2706	6.7732E-03
YJR025C	Bna1p	0.2705	3.2426E-03
YLL024C	Ssa2p	0.2697	1.3281E-03
YER074W-A	Yos1p	0.2684	2.9670E-03
YGL001C	Erg26p	0.2683	1.9171E-03
YLR330W	Chs5p	0.2678	7.4330E-03
YPR187W	Rpo26p	0.2668	1.4389E-03
YOR357C	Snx3p	0.2668	8.9196E-03
YNR038W	Dbp6p	0.2585	4.5148E-03
YCL017C	Nfs1p	0.2554	4.0513E-04
YLR447C	Vma6p	0.2552	9.4134E-04
YML069W	Pob3p	0.2547	5.1584E-04
YLR314C	Cdc3p	0.2535	1.3632E-04
YNL116W	Dma2p	0.2523	3.0042E-04
YLR027C	Aat2p	0.2523	5.0176E-04
YFR010W	Ubp6p	0.2517	2.1263E-04
YLR285W	Nnt1p	0.2510	1.1486E-03
YNL244C	Sui1p	0.2501	3.9347E-04
YOL039W	Rpp2Ap	0.2498	4.6415E-03
YOR021C	Sfm1p	0.2496	5.5401E-04
YMR039C	Sub1p	0.2495	6.9919E-04
YML125C	Pga3p	0.2486	1.4854E-04
YJR077C	Mir1p	0.2468	5.4557E-04
YOL061W	Prs5p	0.2463	6.9554E-04
YPL091W	Glr1p	0.2458	1.5587E-03
YLR083C	Emp70p	0.2456	1.3878E-03
YDR212W	Tcp1p	0.2440	2.2081E-04
YIR012W	Sqt1p	0.2435	6.6174E-04
YER120W	Scs2p	0.2418	5.0849E-04
YCL033C	Mxr2p	0.2417	3.1234E-04
YNL064C	Ydj1p	0.2405	1.1229E-03
YDR519W	Fpr2p	0.2387	1.8722E-04
YER003C	Pmi40p	0.2377	1.4662E-04
YML022W	Apt1p	0.2368	5.0143E-04
YNR043W	Mvd1p	0.2347	2.4011E-04

YOR074C	Cdc21p	0.2341	5.6077E-04
YPL050C	Mnn9p	0.2338	5.6497E-03
YDL126C	Cdc48p	0.2337	5.2171E-03
YLR351C	Nit3p	0.2332	5.2567E-03
YGR074W	Smd1p	0.2318	4.7571E-03
YGL245W	Gus1p	0.2305	2.2314E-04
YLR199C	Pba1p	0.2303	1.5079E-03
YGR207C	Cir1p	0.2301	1.1679E-03
YNR017W	Tim23p	0.2296	4.2834E-03
YDL143W	Cct4p	0.2280	2.3901E-04
YLR089C	Alt1p	0.2279	4.4120E-03
YNL037C	Idh1p	0.2277	4.4900E-03
YER088C	Dot6p	0.2272	1.7991E-03
YKL016C	Atp7p	0.2261	7.1334E-04
YGL105W	Arc1p	0.2253	2.2692E-04
YBR247C	Enp1p	0.2232	1.0758E-03
YOL038W	Pre6p	0.2229	1.9890E-04
YMR246W	Faa4p	0.2228	3.7776E-03
YPL010W	Ret3p	0.2218	7.8561E-04
YPR097W		0.2206	3.5980E-03
YML010W	Spt5p	0.2195	2.0037E-04
YMR186W	Hsc82p	0.2191	8.1202E-04
YPL271W	Atp15p	0.2186	2.7147E-04
YJL174W	Kre9p	0.2186	6.6495E-03
YDL195W	Sec31p	0.2148	1.9815E-03
YDR194C	Mss116p	0.2145	8.5644E-03
YJL173C	Rfa3p	0.2100	4.1753E-03
YNL084C	End3p	0.2079	9.7369E-04
YGR261C	Apl6p	0.2078	4.7366E-03
YLR275W	Smd2p	0.2043	7.3324E-03
YOR027W	Sti1p	0.2039	1.0488E-03
YOR142W	Lsc1p	0.2030	9.0722E-03
YOL064C	Met22p	0.2027	1.9243E-03
YLR397C	Afg2p	0.2025	6.3052E-03
YJL060W	Bna3p	0.2023	4.0549E-03
YBR039W	Atp3p	0.2007	2.1212E-03
YJR045C	Ssc1p	0.1995	9.0464E-04
YGL011C	Scl1p	0.1985	3.5136E-04
YER069W	Arg5,6p	0.1981	5.2131E-03

YNL157W	Igo1p	0.1963	5.8500E-03
YDL132W	Cdc53p	0.1957	7.3490E-03
YDL086W		0.1940	6.1412E-03
YDR299W	Bfr2p	0.1898	5.7126E-03
YBL030C	Pet9p	0.1888	3.6846E-03
YGR175C	Erg1p	0.1878	5.2665E-03
YNL044W	Yip3p	0.1872	2.5168E-03
YPR163C	Tif3p	0.1871	3.0757E-03
YML124C	Tub3p	0.1864	4.2468E-03
YOL016C	Cmk2p	0.1858	1.2289E-03
YOL123W	Hrp1p	0.1808	6.8345E-04
YBR164C	Arl1p	0.1798	9.6582E-03
YOR168W	Gln4p	0.1792	1.1267E-03
YBR198C	Taf5p	0.1783	1.0880E-03
YPL218W	Sar1p	0.1769	1.9987E-03
YDR071C	Paa1p	0.1764	7.8561E-04
YOR046C	Dbp5p	0.1763	1.0084E-03
YLR295C	Atp14p	0.1749	4.5258E-03
YJL159W	Hsp150p	0.1742	6.2778E-03
YKR043C	Shb17p	0.1733	6.6077E-03
YOL097C	Wrs1p	0.1723	4.4346E-03
YDL061C	Rps29Bp	0.1714	5.5340E-03
YDR298C	Atp5p	0.1703	3.0757E-03
YLR180W	Sam1p	0.1702	2.1890E-03
YCR053W	Thr4p	0.1694	5.0186E-03
YDL097C	Rpn6p	0.1685	3.8971E-03
YDR378C	Lsm6p	0.1685	9.8172E-04
YBR248C	His7p	0.1677	3.8758E-03
YGL097W	Srm1p	0.1662	2.1583E-03
YDR028C	Reg1p	0.1646	6.5700E-03
YNL075W	Imp4p	0.1635	2.3736E-03
YJR121W	Atp2p	0.1609	7.4605E-03
YML078W	Cpr3p	0.1576	2.1108E-03
YOR230W	Wtm1p	0.1556	1.9342E-03
YJR017C	Ess1p	0.1526	4.9281E-03
YPL078C	Atp4p	0.1523	3.0381E-03
YDR321W	Asp1p	0.1509	2.9560E-03
YOR065W	Cyt1p	0.1492	2.8791E-03
YOR020C	Hsp10p	0.1478	1.6408E-03

YML028W	Tsa1p	0.1458	1.1903E-03
YFR052W	Rpn12p	0.1455	8.6548E-03
YGR135W	Pre9p	0.1432	5.0488E-03
YMR125W	Sto1p	0.1429	8.4474E-03
YDR174W	Hmo1p	0.1399	4.7366E-03
YLR324W	Pex30p	0.1375	6.0232E-03
YFR016C		0.1346	4.8806E-03
YPL028W	Erg10p	0.1209	4.7415E-03
YPR033C	Hts1p	0.1033	5.6661E-03